Multi-connectivity and resource allocation for slices in 5G networks

Abdellatif Chagdali

université
PARIS-SACLAY

Multi-connectivity and resource allocation for slices in 5G networks

Multi-connectivité et allocation de ressources entre les slices dans la 5G

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Réseaux, information et communications
Graduate School : Sciences de l'ingénierie et des systèmes
Référent : CentraleSupélec

Thèse préparée dans l'unité de recherche Laboratoire des signaux et systèmes (Université Paris-Saclay, CNRS, CentraleSupélec), sous la direction de Salah Eddine Elayoubi, Professeur, et le co-encadrement de Antonia Masucci, ingénieure de recherche

Thèse soutenue à Paris-Saclay, le 21 mars 2022, par

Abdellatif CHAGDALI

Composition du jury

Stefano Secci Professeur, Conservatoire des Arts et Métiers	Président
Urtzi Ayesta Professeur, CNRS/ENSEEIHT Toulouse	Rapporteur & Examinateur
Yezekael Hayel Professeur, Université d'Avignon	Rapporteur & Examinateur
Tijani Chahed Professeur, Télécom SudParis	Examinateur
Philippe Martins Professeur, Télécom Paris	Examinateur
Salah Eddine Elayoubi Professeur, CentraleSupélec	Directeur de thèse
Antonia Masucci Ingénieure de recherche, Orange Labs	Co-encadrente

THESE DE DOCTORAT
NNT : 2022UPAST052

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Multi-connectivité et allocation de ressources entre les slices dans la 5G
Mots clés : Slicing, Théorie des files d'attentes , 5G, URLLC, Multi-connectivité

Résumé : Les futurs réseaux mobiles promettent des opportunités sans précédent pour l'innovation et des cas d'utilisation disruptifs. L'engagement des réseaux 5G et au-delà à fournir des applications critiques nécessite un réseau polyvalent, évolutif, efficace et rentable, capable d'adapter son allocation de ressources pour répondre aux exigences de services hétérogènes. Pour relever ces défis, le découpage du réseau s'est imposé comme l'un des concepts fondamentaux proposés pour améliorer l'efficacité des réseaux mobiles 5G et leur conférer la plasticité requise. L'idée est de fournir des ressources à différentes industries verticales en construisant plusieurs réseaux logiques de bout en bout sur une infrastructure virtualisée partagée. Chaque "tranche de réseau" ainsi définie est personnalisée pour fournir un service spécifique en adaptant son architecture et ses technologies d'accès radio.

Précisément, des applications telles que l'automatisation industrielle ou les communications entre véhicules imposent aux réseaux cellulaires des exigences strictes en matière de latence et de fiabilité. Étant donné que le réseau mobile actuel ne peut pas répondre à ces exigences, les communications ultra-fiables et à faible temps de latence constituent un sujet de recherche essentiel qui a suscité un élan considérable de la part du monde universitaire et des alliances industrielles. Pour répondre à ces exigences, l'utilisation de la multi-connectivité, c'est-à-dire l'exploitation simultanée de plusieurs liaisons radio comme voies de communication, est une approche prometteuse.

L'objectif du présent manuscrit est d'étudier des techniques d'allocation de resources exploitant la couverture redondante des utilisateurs, garantie dans de nombreux scénarios 5G. Nous examinons d'abord l'évolution des réseaux mobiles et discutons des diverses considérations relatives à l'architecture de découpage du réseau et de son impact sur la conception des méthodes d'allocation des ressources. Nous utilisons ensuite les outils de la théorie des files d'attente pour modéliser un système dans lequel un ensemble d'utilisateurs URLLC sont connectés simultanément à deux stations de base ayant la même bande passante ; nous appelons ce scénario le cas homogène. Nous introduisons des politiques d'allocation appropriées et évaluons leurs performances respectives en évaluant leur fiabilité. Ensuite, nous étendons les résultats du cas homogène à un cadre plus général où les interfaces physiques gèrent des bandes passantes différentes, que nous appelons le cas hétérogène. Enfin, nous fusionnons les éléments ci-dessus pour valider le choix des schémas d'allocation des ressources en tenant compte de l'architecture déployée.

Title : Multi-connectivity and resource allocation for slices in 5G networks

Keywords : Network Slicing, Queueing theory, Multi-connectivity, 5G, ultra-reliable low-latency communications

Abstract : Future mobile networks envision unprecedented innovation opportunities and disruptive use cases. As a matter of fact, the 5G and beyond networks' pledge to deliver mission-critical applications mandates a versatile, scalable, efficient, and cost-effective network capable of accommodating its resource allocation to meet the services' heterogeneous requirements. To face these challenges, network slicing has emerged as one of the fundamental concepts proposed to raise the 5G mobile networks' efficiency and provide the required plasticity. The idea is to provide resources for different vertical industries by building multiple end-to-end logical networks over a shared virtualized infrastructure. Each network slice is customized to deliver a specific service and adapts its architecture and radio access technologies.

Precisely, applications such as industrial automation or vehicular communications pose stringent latency and reliability requirements on cellular networks. Given that the current mobile network cannot meet these requirements, ultra-reliable low latency (URLLC) communications embodies a vital research topic that has gathered substantial momentum from academia and industrial alliances. To reach URLLC requirements, employing multi-connectivity, i.e., exploiting multiple radio links as communication paths at once, is a promising approach.

Therefore, the objective of the present manuscript is to investigate dynamic scheduling techniques, exploiting redundant coverage of users guaranteed in numerous 5G radio access network scenarios. We first review the evolution of mobile networks and discuss various considerations for network slicing architecture and its impact on resource allocation design. Then, we use tools from queuing theory to model a system in which a set of URLLC users are connected simultaneously to two base stations having the same bandwidth ; we refer to this scenario as the homogenous case. We introduce suitable scheduling policies and evaluate their respective performances by assessing their reliability. Next, we extend the homogenous case's results to a more general setting where the physical interfaces manage different bandwidths, referred to as the heterogeneous case. Finally, we merge the above elements to validate the choice of resource allocation schemes considering the deployed architecture.

DOCTORAL THESIS
PARIS-SACLAY UNIVERSITY

Speciality: **Networks, information and communications**

Sciences and Technologies of Information and Communication
Doctoral School

Presented by: **Abdellatif Chagdali**

This dissertation is submitted for the degree of

Doctor of Philosophy

Multi-connectivity and resource allocation for slices in 5G networks

---

Committee:

Urtzi Ayesta, IRIT/ENSEEIHT Toulouse	Reviewer
Yezekael HAYEL, Avignon University	Reviewer
Tijani CHAHED, Télécom SudParis	Examiner
Philippe MARTINS, Télécom ParisTech	Examiner
Stefano SECCI, Conservatoire National des Arts et Métiers	Examiner
Salah Eddine ELAYOUBI, CentraleSupélec	Doctoral supervisor
Antonia MASUCCI, Orange Labs	Advisor



THÈSE DE DOCTORAT
UNIVERSITÉ PARIS-SACLAY

Spécialité: **Réseaux, information et communications**

École doctorale sciences et technologies de l'information
et de la communication (ED STIC)

Présentée par: **Abdellatif Chagdali**

en vue de l'obtention du grade de

Docteur de L'université Paris-Saclay

Multi-connectivité et allocation de ressources entre les slices dans la 5G

Jury:

Urtzi AYESTA, IRIT/ENSEEIHT Toulouse	Rapporteur
Yezekael HAYEL, Université d'Avignon	Rapporteur
Tijani CHAHED, Télécom SudParis	Examinateur
Philippe MARTINS, Télécom ParisTech	Examinateur
Stefano SECCI, Conservatoire National des Arts et Métiers	Examinateur
Salah Eddine ELAYOUBI, CentraleSupélec	Directeur de thèse
Antonia MASUCCI, Orange Labs	Encadrante

RÉSUMÉ

Les futurs réseaux mobiles promettent des opportunités sans précédent pour l'innovation et des cas d'utilisation disruptifs. L'engagement des réseaux 5G et au-delà à fournir des applications critiques nécessite un réseau polyvalent, évolutif, efficace et rentable, capable d'adapter son allocation de ressources pour répondre aux exigences de services hétérogènes. Pour relever ces défis, le découpage du réseau s'est imposé comme l'un des concepts fondamentaux proposés pour améliorer l'efficacité des réseaux mobiles 5G et leur conférer la plasticité requise. L'idée est de fournir des ressources à différentes industries verticales en construisant plusieurs réseaux logiques de bout en bout sur une infrastructure virtualisée partagée. Chaque "tranche de réseau" ainsi définie est personnalisée pour fournir un service spécifique en adaptant son architecture et ses technologies d'accès radio.

Précisément, des applications telles que l'automatisation industrielle ou les communications entre véhicules imposent aux réseaux cellulaires des exigences strictes en matière de latence et de fiabilité. Étant donné que le réseau mobile actuel ne peut pas répondre à ces exigences, les communications ultra-fiables et à faible temps de latence constituent un sujet de recherche essentiel qui a suscité un élan considérable de la part du monde universitaire et des alliances industrielles. Pour répondre à ces exigences, l'utilisation de la multi-connectivité, c'est-à-dire l'exploitation simultanée de plusieurs liaisons radio comme voies de communication, est une approche prometteuse.

L'objectif du présent manuscrit est d'étudier des techniques d'allocation de

resources exploitant la couverture redondante des utilisateurs, garantie dans de nombreux scénarios 5G. Nous examinons d'abord l'évolution des réseaux mobiles et discutons des diverses considérations relatives à l'architecture de découpage du réseau et de son impact sur la conception des méthodes d'allocation des ressources. Nous utilisons ensuite les outils de la théorie des files d'attente pour modéliser un système dans lequel des utilisateurs URLLC sont connectés simultanément à deux stations de base ayant la même bande passante ; nous appelons ce scénario le cas homogène. Nous introduisons des politiques d'allocation appropriées et évaluons leurs performances respectives en évaluant leur fiabilité. Ensuite, nous étendons les résultats du cas homogène à un cadre plus général où les interfaces physiques gèrent des bandes passantes différentes, que nous appelons le cas hétérogène. Enfin, nous fusionnons les éléments ci-dessus pour valider le choix des schémas d'allocation des ressources en tenant compte de l'architecture déployée.

ABSTRACT

Future mobile networks envision unprecedented innovation opportunities and disruptive use cases. As a matter of fact, the 5G and beyond networks' pledge to deliver mission-critical applications mandates a versatile, scalable, efficient, and cost-effective network capable of accommodating its resource allocation to meet the services' heterogeneous requirements. To face these challenges, network slicing has emerged as one of the fundamental concepts proposed to raise the 5G mobile networks' efficiency and provide the required plasticity. The idea is to provide resources for different vertical industries by building multiple end-to-end logical networks over a shared virtualized infrastructure. Each network slice is customized to deliver a specific service and adapts its architecture and radio access technologies.

Precisely, applications such as industrial automation or vehicular communications pose stringent latency and reliability requirements on cellular networks. Given that the current mobile network cannot meet these requirements, Ultra-Reliable Low-Latency Communication (URLLC) communications embodies a vital research topic that has gathered substantial momentum from academia and industrial alliances. To reach URLLC requirements, employing Multi-Connectivity (MC), i.e., exploiting multiple radio links as communication paths at once, is a promising approach.

The objective of the present manuscript is to investigate dynamic scheduling techniques, exploiting redundant coverage of users guaranteed in numerous 5G radio access network scenarios. We first review the evolution of mobile networks and discuss various considerations for network slicing architecture

and its impact on resource allocation design. Then, we use tools from queuing theory to model a system in which a set of URLLC users are connected simultaneously to two base stations having the same bandwidth; we refer to this scenario as the homogenous case. We introduce suitable scheduling policies and evaluate their respective performances by assessing their reliability. Next, we extend the homogenous case's results to a more general setting where the physical interfaces manage different bandwidths, referred to as the heterogeneous case. Finally, we merge the above elements to validate the choice of resource allocation schemes considering the deployed architecture.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors, Prof. Salah Eddine Elayoubi and Dr. Antonia Maria Masucci. Their dedication, commitment, and love for their work have always motivated me. Their valuable advice and mentoring have set me on the right track to present this work. Their unmatched devotion to their daily tasks is proof of boundless esteem for their respective teams and colleagues at CentraleSupélec and Orange Labs. In the words of Simone Weil, "Attention is the rarest and purest form of generosity." so I would like to hereby state my gratefulness for the attention they paid to every detail. Working with them for the last three years was an honor and an informative experience. Even though a global pandemic has stalled our work rhythm, they remained supportive and adapted to this unprecedented situation for which we all encumbered the ramifications, one way or another.

I would also like to thank the members of the "Ph.D. group" with whom I enjoyed working every day: Adrien Cambier, Ayman Chouayakh, Wesley Coelho, Raphael Colares, Thibaut Cuvelier, Chen Dang, Quentin le Gall, Ahlam Mouaci, Georges Nassif, Alexandre Pacaud, Marine Picot, Raquel Rugani Lage, and Paolo Zappala. The environment at Orange Labs has contributed immensely to appreciating my work. A special thank must go to Alain Simonian, my mentor, who, in addition to being a virtuoso musician, is also a very knowledgeable person. I had a lot of pleasure discussing history, cinema, and literature with him. I would like to thank My colleagues at Orange Labs, with whom I had the excellent chance to discuss, debate, and laugh... My integration into the team was easy because I had the luck

to meet such welcoming and easy-going people. Special thanks go to my manager at Orage Labs, Eric Gourdin, for his infallible support and dedication to all administrative drudgery, with which he dealt with a lot of humor and openmindedness. Those moments will permanently be engraved in my memory.

And since I'm aware that my words can never surpass those of Anais Nin, who said, "Each friend represents a world in us, a world possibly not born until they arrive, and it is only by this meeting that a new world is born." I would like to express, on a personal note, my gratitude to Halima Derbani, Ibtissam Labriji, Omayma Ghitou, Paloma Ruiz, Pauline Huau, and Zaid Hammouch for their friendship. For once, I'll allow myself to be sentimental and say that my life is a lot better with you in it. I must thank my parents for supporting me throughout the years and for being my most ardent fans. Words fail to express how much I love you, so I would let the silence reign, for what good can ill-written sentences and half-kept promises bring.

Each generation doubtless feels called upon to reform the world. Mine knows that it will not reform it, but its task is perhaps even greater. It consists in preventing the world from destroying itself.

Albert Camus, *Nobel prize speech*

Anyone whose goal is 'something higher' must expect someday to suffer vertigo.

Milan Kundera, *The Unbearable Lightness of Being*

CONTENTS

Résumé **i**

Abstract **iii**

Acknowledgements **v**

List of Figures **xiii**

List of Tables **xvi**

Introduction générale **xxi**

1 Introduction **1**

1.1 Background and general context 1

1.2 Motivation 3

1.3 Thesis Outline and document structure 4

2 Evolution of mobile networks towards network slicing 7

2.1 From distributed to centralized radio access networks 8

2.2 From 4G C-RAN to 5G 10

2.3 From monolithic to slicing-empowered systems 12

2.4 5G ecosystem: business relationships and service level agreements 14

2.4.1 Business relationships between actors 14

2.4.2 Service Level Agreements 15

2.5 Summary 17

3 Network Slicing architecture 18

3.1 Slice management functions description 18

3.1.1 Resource allocation from MSP perspective 20

3.1.2 Resource allocation from tenant perspective 21

3.1.3 Resource allocation from InP perspective 21

3.2 Impact of Placement of intelligent entities 22

3.2.1 Intelligence placed at the level of a shared RAN NSSMF 23

3.2.2 Intelligence placed at the NSMF level 24

3.3 Summary 25

4 Multi-connectivity for URLLC slice in homogeneous settings 26

4.1 System Model 27

4.2 Main results 29

4.2.1 Occupancy distributions 29

4.2.2 Decay Rates 32

4.2.3 Outage Probability 38

4.2.4 Average Sojourn Time 43

4.3 Resource dimensioning 44

4.4 Summary 46

5 Multi-connectivity for URLLC slices: Extension to the heterogeneous case 47

5.1 System Model 48

5.1.1 Resource allocation schemes 48

5.1.2 Queuing model 48

5.2 Performance Evaluation 49

5.2.1 Delay distribution of the JSQ scheduling scheme . . . 50

5.2.2 Delay distribution of the SED system 53

5.2.3 Delay distribution of the RED system 54

5.2.4 Delay distribution of the CAN system 57
5.3 Numerical experiments 58
5.3.1 Qualitative Analysis 59
5.4 Summary 61

6 Assessing architecture impact on scheduling policies 62
6.1 Simulations for systems serving URLLC slices 63
6.1.1 System model 63
6.2 Performance evaluation 65
6.2.1 Bandwidth Reservation Case for URLLC Slice 65
6.2.2 Coexistence of eMBB and URLLC Slices 67
6.3 Case Study: a smart factory served by three BSs 70
6.4 Summary 71

7 Conclusion and perspectives 73
7.1 Concluding Remarks 73
7.2 Perspectives 75

Appendix A Large deviation theory 77

Appendix B Industry 4.0 slicing use case illustration 81
B.1 Use case description 81

B.1.1 mMTC traffic 82

B.1.2 eMBB traffic 83

B.1.3 URLLC traffic 83

B.2 Slice management for the factory 84

Publications **86**

Bibliography **87**

LIST OF FIGURES

2.1 Evolution of cellular networks from distributed RAN to centralized RAN (adapted from [1]). 9

2.2 Functional split options between centralized, distributed and radio units, as defined by 3GPP [2]. 11

2.3 Evolution of cellular networks from centralized RAN to NFV/ SDN empowered 5G networks (adapted from [1]). 12

2.4 Evolution of cellular networks from NFV/SDN to network slicing based network (adapted from [1]). 13

2.5 Business relationships in the 5G RAN. 16

3.1 Overview of the infrastructure and management Layer in a network slicing scenario. 20

3.2 Traffic steering in industry 4.0 use case. 22

3.3 Intelligence placed at the level of a shared RAN NSSMF. . . . 23

3.4 Intelligence placed at the NSMF level. 24

4.1 Two BSs in the neighborhood of URLLC user equipment. . . 27

4.2 Respective decay rates S^*, Σ^* and Ξ^* for JSQ, RED and CAN allocation schemes. 38

4.3 Mean sojourn time $\mathbb{E}[T]$ for JSQ, RED and CAN in terms of ϱ. 44

4.4 Outage probability $\mathbb{P}(T > t_0)$ with $t_0 = 1$ ms for JSQ, RED and CAN in terms of ϱ. 45

4.5 Maximum achievable load and arrival rate for JSQ, RED and CAN. 46

5.1 Outage probability for the JSQ scheme using equilibrium equations and simulations with $t_0 = 0.5$ ms 53

5.2 Outage probability for the SED scheme using simulations with $t_0 = 0.5$ ms 54

5.3 Outage probability comparison of the JSQ and SED schemes using simulations with $t_0 = 0.5$ ms 55

5.4 Outage probability for the RED scheme using equilibrium equations and simulations with $t_0 = 0.5$ ms 56

5.5 Outage probability $\mathbb{P}(T > t_0)$ with $t_0 = 0.5$ ms for CAN in terms of λ for different values of z. 58

5.6 Outage probability $\mathbb{P}(T > t_0)$ with $t_0 = 0.5$ ms for JSQ, SED, RED and CAN in terms of λ for different values of z. 59

6.1 Outage probability in the case of bandwidth reservation for URLLC packets. **(a)** $B_1 = B_2 = 1$ MHz; **(b)** $B_1 = 2, B_2 = 1$ MHz 66

6.2 Outage probability in the case of fixed eMBB and variable URLLC traffic without bandwidth reservation.**(a)** $B_1 = B_2 =$ 10 MHz; **(b)** $B_1 = 20, B_2 = 10$ MHz 67

6.3 Number of packets served in BS1, BS2 and the system, respectively with $B_1 = B_2 = 10$ MHz.**(a)** BS1 ; **(b)** BS2 ; **(c)** System 68

6.4 Outage probability in the case of variable eMBB at BS1, fixed eMBB traffic at BS2, and fix number of URLLC users.**(a)** $B_1 = B_2 = 10$ MHz; **(b)** $B_1 = 20, B_2 = 10$ MHz 69

6.5 eMBB throughput in the case of variable eMBB traffic at BS1, fixed eMBB traffic at BS2, and fix number of URLLC users, where $B_1 = B_2 = 10$ MHz. 70

6.6 Outage probability in the case of bandwidth reservation for URLLC packets. **(a)** $B_1 = B_2 = B_3 = 1$ MHz; **(b)** $B_1 = 2, B_2 = B_3 = 1$ MHz 71

B.1 Industry 4.0 slices from the tenant perspective: the URLLC slice is locally based while the mMTC and eMBB slices reach external networks to the factories premise. 82

LIST OF TABLES

3.1 Entities involved in RAN resource allocation and their roles. . 19

6.1 Parameters for performance evaluation 65

ACRONYMS

2G 2nd generation of mobile networks.

3GPP 3rd Generation Partnership Project.

4G 4th generation of mobile networks.

5G 5th generation of mobile networks.

AU Access Unit.

BBU Baseband Unit.

BS Base Station.

C-RAN Centralized RAN.

CAN Redundancy with Cancellation.

CN Core Network.

CPRI Common Public Radio Interface.

CSMF Communication Service Management Function.

CU Centralized Unit.

D-RAN Distributed RAN.

DU Distributed Unit.

DU Digital Unit.

EB Exabytes.

eCPRI enhanced Common Public Radio Interface.

eMBB enhanced Mobile Broadband.

eNodeB Evolved Node B.

EPC Evolved Packet Core.

GSM Global System for Mobile Communications.

InP Infrastructure Provider.

IoT Internet of Things.

JSQ Join-the-Shortest-Queue.

KPI Key Performance Indicator.

LTE Long Term Evolution.

M2M Machine-to-Machine.

MANO Management and orchestration.

MC Multi-Connectivity.

MCS Modulation and Coding Scheme.

MEC Mobile Edge Computing.

MIMO Massive multiple-input multiple-output.

mMTC massive Machine Type Communications.

MSP Mobile Service Provider.

MVNO Mobile Virtual Network Operators.

N3IWF Non-3GPP Interworking Function.

NF Network Function.

NFV Network Function Virtualization.

NSI Network Slice Instance.

NSMF Network Slice Management Function.

NSSI Network Slice Subnet Instance.

NSSMF Network Slice Subnet Management Function.

OFDM Orthogonal Frequency-Division Multiplexing.

OTT Over-The-Top.

PDCP Packet Data Convergence Protocol.

PHY Physical Layer.

PNF Physical Network Function.

QoS Quality of Service.

RAN Radio Access Network.

RANaaS RAN-as-a-Service.

RATs Radio Access Technologies.

RED Redundancy.

RF Radio Frequency.

RL reinforcement Learning.

RLC Radio Link Control.

RRH Remote Radio Head.

RRM Radio Resource Management.

RU Radio Unit.

SCMA Sparse Code Multiple Access.

SDN Software-Defined Networking.

SED Shortest Expected Delay.

SLA Service Level Agreement.

TTI Transmission Time Interval.

UE User Equipment.

UMTS Universal Terrestrial Radio Telecommunication System.

URLLC Ultra-Reliable Low-Latency Communication.

VNF Virtual Network Function.

vRAN virtual RAN.

INTRODUCTION GÉNÉRALE

Historique et contexte général

L'odyssée des télécommunications a commencé le 10 mars 1876, lorsque Alexander Graham Bell a passé le premier appel à son assistant, en prononçant les mots suivants : *"M. Watson, venez ici. Je veux vous voir."* [3], sans se douter des conséquences sociologiques, politiques, environnementales, philosophiques et même esthétiques d'une telle ampleur, au point de modifier notre perception du monde. Près d'un siècle plus tard, le 3 avril 1973, Martin Cooper, chercheur chez Motorola, a passé le premier appel téléphonique mobile à partir d'un téléphone portable de poche. Il tentait de faire une démonstration de son prototype à un journaliste lorsqu'il a eu l'idée d'appeler le siège de ses concurrents, Bell Labs, dans le New Jersey. Il se trouvait près d'une station de base de 900 MHz lorsqu'il a appelé le siège de son rival, le Dr Joel S. Engel, pour l'informer de la percée qu'il avait réalisée en utilisant l'un des deux exemplaires du prototype légendaire. Cet événement a progressivement placé les dispositifs de communication sans fil à l'épicentre des interactions humaines, modifiant ainsi radicalement le paysage social et économique. Quarante-huit ans plus tard, nous vivons sur une planète qui compte beaucoup plus d'appareils que d'êtres humains.

Une autre étape importante dans l'histoire des télécommunications a été franchie le 1er juillet 1991, quand Harri Holkeri, alors premier ministre de la Finlande, a passé le premier appel mobile numérique au monde par le biais du Groupe spécial mobile (abrégé en GSM, sigle de l'anglais Global System for Mobile Communications) à Helsinki [4]. Cette technologie est également connue sous le nom de la deuxième génération de réseaux mobiles (2G), faisant suite à la première génération mise en œuvre dans les années 1980, basée sur des technologies analogiques. Ces deux générations étaient principalement utilisées pour les applications vocales et étaient exclusivement à commutation de circuits.

Au cours des 30 dernières années, on a assisté à une croissance exponentielle des services à haut débit d'accès, ce qui a entraîné l'inauguration d'une nouvelle génération tous les dix ans. Chaque nouvelle génération, jusqu'à présent, a garanti des débits de données plus élevés tout en donnant naissance à de nouvelles applications et de nouveaux marchés. Par exemple, GSM offrait aux utilisateurs des débits de données de quelques dizaines de kilobits par seconde. La troisième génération, appelée Universal Terrestrial Radio Telecommunication System (UMTS), a fourni pour la première fois un accès Internet à haut débit et des expériences mobiles à large bande améliorées.

Néanmoins, le besoin continu et croissant de débits de données plus élevés a conduit au déploiement de la quatrième génération des réseaux mobiles appelée Long Term Evolution (LTE). L'avènement de la technologie LTE a entraîné une croissance massive du trafic de données mobiles. Cette incontestable augmentation est liée à l'utilisation généralisée d'appareils centrés sur les données comme les smartphones et les tablettes. En fait, le nombre d'appareils connectés à internet (il s'agit d'Internet des objets, abrégé en IoT, sigle de l'anglais Internet of Things) dans le monde en 2020 est d'environ 11,5 milliards et devrait passer à 38,6 milliards en 2025, selon les estimations.

Les générations précédentes de réseaux cellulaires se sont concentrées sur les cas d'utilisation centrés sur l'homme, marquant des améliorations dans les capacités atteintes et les débits de données. La cinquième génération des réseaux mobiles 5G diffère de ses prédécesseurs car elle abrite une vision orientée vers des nouveaux services et sans précédent, ainsi qu'une vision évolutive. D'une part, la 5G offrira des services à haut débit mobile amélioré (abrégé eMBB, sigle de l'anglais enhanced Mobile Broadband) qui nécessitent une couverture radio sans faille, offrant ainsi des débits de données

expérimentés élevés et ubiquitaires. Parmi les cas d'utilisation envisagés, nous mentionnons la réalité augmentée, l'expérience d'événements immersifs et les vidéos 8K. D'autre part, les systèmes 5G vont également initier une évolution disruptive, définissant de nouveaux cas d'utilisation et services en plus des services existants. Ces nouveaux services peuvent être séparés en deux catégories. Tout d'abord, la catégorie des communications de type machine massive (abrégé mMTC, sigle de massive Machine Type Communications), prend en compte le déploiement d'un large ensemble de dispositifs ayant des exigences disparates en termes de débit de données, de fiabilité et de latence, entraînant une augmentation exponentielle du nombre de dispositifs connectés [5, 6]. Deuxièmement, la famille de services liée aux communications à faible latence très fiables(abrégé en URLLC, sigle de l'anglais Ultra-Reliable Low-Latency Communication) définit des exigences strictes en termes de latence et de perte de paquets pour des applications dans les secteurs du médical et des voitures autonomes, par exemple [7].

Motivation

La provision de ces nouveaux services nécessite un réseau polyvalent, évolutif, efficace et rentable, capable de s'adapter à l'allocation de ses ressources pour agir sur la nature hétérogène, onéreuse et parfois conflictuelle des demandes. L'approche classique consistant à déployer un réseau monolithique pour répondre aux demandes de diverses industries verticales avec des exigences variées en matière de qualité de service (abrégé en QoS, sigle de l'anglais Quality of Service) est obsolète, car elle augmente les dépenses d'investissement et les dépenses opérationnelles tout en entraînant une sous-utilisation des ressources. Compte tenu de ces points, le découpage du réseau (en anglais, network slicing) est apparu comme l'un des concepts fondamentaux proposés pour améliorer l'efficacité et fournir la plasticité requise des réseaux mobiles 5G [8, 9]. L'idée est de fournir des ressources à différentes industries verticales en construisant plusieurs réseaux logiques de bout-en-bout (abrégé E2E, sigle de l'anglais end-to-end) sur une infrastructure partagée. Chaque tranche de réseau, c'est-à-dire chaque réseau logique, est conçue pour fournir un service spécifique à un vertical.

Bien que le concept de découpage du réseau soit relativement nouveau, la littérature correspondante qui en traite est déjà abondante, notamment sur les

aspects d'architecture et de gestion. Par exemple, [10] propose une approche holistique en discutant de la gestion et de l'orchestration pour les tranches de bout-en-bout, notamment la couche infrastructure, la couche fonction réseau et la couche service. [11] aborde les concepts architecturaux du découpage en tranches, notamment le chaînage des fonctions réseau (abrégé NF, sigle de l'anglais Network Function) pour satisfaire les objectifs de performance hétérogènes. Bien que le découpage du réseau soit un concept de bout-en-bout, la plupart des recherches se sont concentrées sur le découpage du cœur de réseau, ce qui a conduit à des propositions d'architecture matures alimentées par l'émergence du cloud computing, la virtualisation des fonctions réseau (abrégé NFV, sigle de l'anglais Network Function Virtualization), les réseaux SDN (Software-Defined Networking, en anglais) [10, 12, 13].

Cependant, le découpage du réseau d'accès radio (abrégé RAN, sigle de l'anglais Radio Access Network) pose des problèmes distincts de ceux du noyau. Le projet de partenariat de la 3rd Generation Partnership Project (3GPP) prévoit de nouvelles technologies d'accès radio (abrégé RAT, de l'anglais Radio Access Technologies), un nouvel espacement des sous-porteuses et une nouvelle structure de trame pour assurer l'adaptabilité du RAN, étant donné la nature discordante des demandes verticales [14]. Si de nombreux travaux dans la littérature ont ouvert la voie à la définition du concept de découpage en tranches dans la 5G, ils n'ont pas abordé l'aspect critique de la mise en œuvre de l'allocation des ressources dans le RAN. En effet, même si la nouvelle radio (NR) 5G a été conçue comme très flexible pour assurer un multiplexage efficace entre les tranches du réseau, la tâche d'allocation des ressources radio et informatiques reste pesante. La multiplication des acteurs ayant des intérêts dans le RAN rend difficile l'allocation des ressources aux tranches, car les ressources sont censées appartenir à plusieurs fournisseurs d'infrastructures (abrégé InP, sigle de l'anglais Infrastructure Provider). Ces derniers concluent des accords de niveau de service (abrégé SLA, sigle de l'anglais Service Level Agreement) avec différents fournisseurs de services mobiles (abrégé MSP, sigle de l'anglais Mobile Service Provider) et verticaux. De plus, la fourniture de services URLLC, en particulier, va engendrer de nombreux défis sur le réseau actuel centré sur la capacité en raison des contraintes strictes de latence et de fiabilité nécessaires.

Par conséquent, cette thèse se concentre sur la description de l'écosystème complexe de la 5G et, à sa lumière, sur le développement de schémas d'ordonnancement de paquets pour URLLC exploitant la multi-connectivité; ce qui

signifie que les utilisateurs se connectent au réseau via plusieurs chemins en s'attachant à l'intégration de plusieurs interfaces physiques au sein du RAN 5G [15]. Lorsque la quantité de ressources réservées à l'URLLC est limitée, l'ordonnancement redondant sur plusieurs ressources disponibles est un moyen pratique de réduire les délais d'attente puisque les systèmes 5G sont conçus pour combiner plusieurs RAT, y compris la 5G NR avec plusieurs bandes de fréquences et interfaces. En pratique, la couverture redondante est assurée dans presque tous les endroits, notamment dans les zones denses [16]. Par conséquent, l'exploitation de la diversité spatiale, c'est-à-dire le fait que plusieurs stations de base couvrent le même dispositif, est un moyen de réduire les délais d'ordonnancement et de mise en file d'attente en sélectionnant dynamiquement la station de base ayant la plus petite charge instantanée ou en répliquant le paquet sur plusieurs stations de base.

Structure de la thèse

Ce manuscrit est organisé comme suit. Nous présentons d'abord brièvement l'évolution des réseaux mobiles d'une architecture distribuée à une architecture centralisée et finalement à l'avènement du découpage de réseau en tant que facilitateur des systèmes 5G dans le chapitre 2. Nous présentons les relations commerciales entre les différents acteurs de l'écosystème 5G et les accords de niveau de service entre eux.

Dans le chapitre 3, nous identifions le rôle de chaque acteur dans la gestion des tranches de réseau et les entités chargées d'accueillir le trafic et les ressources pour les tranches. Ensuite, nous décrivons les options de placement de la fonction de gestion des tranches avant de nous pencher sur les défis des services URLLC et sur un ensemble de facilitateurs pour ces cas d'utilisation de pointe. Une partie du contenu présenté dans les chapitres 2 et 3 a été publiée dans le compte rendu de la 16e conférence internationale sur l'informatique, les réseaux et les communications sans fil et mobiles (WiMob) [17].

Dans le chapitre 4, nous étudions la performance des schémas d'ordonnancement des paquets pour les services URLLC. Afin de garantir un faible temps d'attente, nous exploitons la couverture redondante dans de nombreux scénarios de déploiement du réseau d'accès radio dans la 5G, où deux tech-

nologies d'accès radio ayant le même temps de service sont intégrées. Nous considérons trois schémas d'ordonnancement et de redondance des paquets, à savoir rejoindre la file d'attente la plus courte, Redondance systématique et Redondance avec Annulation à l'achèvement (abrégé JSQ, RED et CAN, sigle de l'anglais Join-the-Shortest-Queue, Redundancy, et Redundancy with Cancellation, respectivement). Sur la base d'une analyse de la théorie des files d'attente, nous développons des expressions pour la fiabilité, définie comme la probabilité que le paquet soit transmis avant un délai cible donné. Nous montrons que RED est performant à faible charge, tandis que JSQ est meilleur lorsque la charge augmente ; CAN surpasse tous les autres schémas. Nous montrons ensuite comment les résultats obtenus peuvent être utilisés pour dimensionner les ressources radio nécessaires à la 5G et discutons du compromis entre performance et complexité de mise en œuvre. Le contenu du chapitre 4 a été présenté au Congrès international sur le télétrafic (ITC 32) et publié dans ses actes [18].

Dans le chapitre 5, nous exploitons la couverture redondante des utilisateurs pour examiner l'impact de la multi-connectivité dans le cas où les stations de base ont des capacités de service différentes. Comme dans le chapitre 4, nous étudions la performance des schémas d'ordonnancement des paquets pour les services URLLC. Ainsi, nous passons en revue quatre schémas d'ordonnancement et de redondance de paquets, à savoir JSQ, RED, CAN comme mentionné précédemment, et une politique qui prend en considération le plus court retard prévu (abrégé SED, sigle de l'anglais Shortest Expected Delay). Les résultats s'étendent au cas hétérogène où la bande passante réservée dans les deux stations est dissemblable.

Le chapitre 6 de cette thèse évalue l'impact des choix architecturaux discutés dans le chapitre 3 sur la qualité de service de différents services, en se concentrant sur les applications de communication ultra-fiable à faible latence. Nos expériences numériques montrent que le placement des fonctions de gestion des slices joue un rôle crucial dans le choix du schéma d'allocation des ressources radio qui convient le mieux aux slices URLLC dans les cas homogène et hétérogène. Nous étudions également la coexistence du trafic URLLC et eMBB afin de quantifier l'impact de la desserte du premier sur le second. Le contenu présenté dans ce chapitre a été publié dans MDPI computers dans le numéro spécial "Selected Papers from 16th Wireless and Mobile Computing, Networking And Communications (WiMob 2020)". [19].

Le chapitre 7 résume les idées présentées dans cette thèse et propose des pistes de recherche à explorer.

CHAPTER 1

INTRODUCTION

Contents

1.1	**Background and general context**	**1**
1.2	**Motivation**	**3**
1.3	**Thesis Outline and document structure**	**4**

1.1 Background and general context

The telecommunication odyssey began on March 10, 1876, when Alexander Graham Bell made the first call to his assistant, uttering the following words: *"Mr. Watson, come here. I want to see you."* [3], not being aware of the far-reaching sociological, political, environmental, philosophical, and even aesthetic consequences, to the point of actually modifying how we experience the world. Almost one century later, On April 3, 1973, Motorola's researcher Martin Cooper made the first mobile phone call from a portable handheld cellphone. He was trying to demonstrate his prototype to a journalist when he had the idea to call the headquarters of his competitors, Bell Labs in New Jersey. He was standing near a 900 MHz base station when he called his rival's landline, Dr. Joel S. Engel, to inform him of the breakthrough he had made using one of the two copies of the legendary prototype.

This event has gradually put wireless communications devices at the epicenter of human interactions, thus drastically changing the social and economic

landscape. Forty-eight years later, we live on a planet with far more devices than human beings. Another milestone in the history of telecommunications was reached on July 1, 1991, when Harri Holkeri, then prime minister of Finland, made the world's first digital mobile call through the Global System for Mobile Communications (GSM) in Helsinki [4]. This technology is also known as the 2nd generation of mobile networks (2G), following the first generation implemented during the 1980s, based on analog technologies. These two generations were mainly used for voice applications and were exclusively circuit-switched. In the last 30 years, there has been an exponential growth of higher access speed services resulting in the inauguration of a new generation every decade. Every new generation, thus far, has guaranteed higher data rates while birthing new applications and markets. For instance, GSM offered users data rates of tens of kilobits per second. The third generation called the Universal Terrestrial Radio Telecommunication System (UMTS), delivered high-speed internet access and enhanced mobile broadband experiences for the first time.

Nevertheless, the continuous and increasing need for higher data rates has led to the deployment of 4th generation of mobile networks (4G), called Long Term Evolution (LTE) networks. The advent of LTE technology has led to a massive growth of mobile data traffic. This conspicuous increase is linked to the widespread use of data-centric devices as smartphones and tablets. In fact, the number of Internet of Things (IoT) connected devices worldwide in 2020 is around 11.5 billion and is projected to rise to an estimated 38.6 billion by 2025 [20].

Previous generations of cellular networks have focused on human-centric use cases, marking improvements in achieved capacities and data rates. The 5th generation of mobile networks (5G) differs from its predecessors as it harbors a novel and unprecedented service-oriented vision together with the evolutionary view. On the one hand, 5G will offer enhanced Mobile Broadband (eMBB) services that require seamless radio coverage, thereupon offering ubiquitous high experienced data rates. Among the addressed use cases, we mention augmented reality, immersive event experience, and 8K videos [21]. On the other hand, 5G systems will also initiate a disruptive evolution, defining new use cases and services in addition to legacy services. These new services can be separated into two categories. First, massive Machine Type Communications (mMTC) category, which takes into consideration the deployment of a large set of devices with disparate requirements in terms of data

rate, reliability, and latency, leading to an exponential surge in the number of connected devices [5, 6]. Second, the Ultra-Reliable Low-Latency Communication (URLLC) family of services define strict requirements in terms of latency and packet loss for application in the medical and autonomous cars sectors, for example [7].

1.2 Motivation

Delivering these new services requires a versatile, scalable, efficient, and cost-effective network capable of accommodating its allocation of resources to act upon the heterogeneous, onerous, and sometimes conflicting nature of demands. The classical approach of deploying a monolithic network, dealing with various vertical industries demands with varied quality of service (QoS) requirements is obsolete, for it will raise the capital expenditure (CAPEX) and the operational expenditure (OPEX) costs while instigating the under-utilization of resources. Given these points, network slicing has emerged as one of the fundamental concepts proposed to raise the efficiency and provide the required plasticity of 5G mobile networks [8, 9]. The idea is to provide resources for different vertical industries by building multiple end-to-end (E2E) logical networks over a shared infrastructure. Each network slice, meaning each logical network is tailored to deliver a specific service to a tenant[22].

Even though the concept of network slicing is relatively new, the corresponding literature dealing with it is already ample, especially on architecture and management aspects. For instance, [10] offers a holistic approach by discussing the management and orchestration for E2E slices, including infrastructure layer, network function layer, and service layer. [11] discusses the architectural concepts for slicing, including Network Function (NF) chaining for satisfying heterogeneous performance targets. While network slicing is an E2E concept, most of the research has focused on core slicing leading to mature architecture propositions powered by the emergence of cloud computing, Network Function Virtualization (NFV), Software-Defined Networking (SDN) [10, 12, 13].

However, Radio Access Network (RAN) slicing introduces distinct issues compared to core slicing. Third Generation Partnership Project (3GPP)

foresees novel Radio Access Technologies (RATs), new subcarrier spacing, and frame structure to provide RAN adaptability, given the discordant nature of verticals' demands [14].

While numerous works in the literature paved the way for the definition of the slicing concept in 5G, they did not tackle the critical aspect of the implementation of resource allocation in the RAN. Indeed, even if the 5G New Radio (NR) has been designed as highly flexible to ensure efficient multiplexing between slices, the task of radio and computing resource allocation is still cumbersome. The multiplication of actors with stakes in the RAN makes it challenging to allocate resources to the slices, as the resources supposedly belong to multiple Infrastructure Provider (InP). The latter InPs contract Service Level Agreements (SLAs) with different Mobile Service Provider (MSP) and verticals. In addition, providing URLLC services, in particular, will instigate many challenges on the current capacity-centered network because of the stringent latency and reliability constraints needed.

Accordingly, this thesis focuses on describing the complex 5G ecosystem and, in its light, developing packet scheduling schemes for URLLC exploiting multi-connectivity; meaning that users connect to the network via multiple paths by attaching to the integration of several Physical Layer (PHY) interfaces within the 5G RAN [15]. When the amount of resources reserved for URLLC is limited, redundant scheduling over several available resources is a practical way for reducing queuing delays since 5G systems are designed to combine multiple RATs, including 5G NR with several frequency bands and interfaces. In practice, redundant coverage is ensured in almost all locations, especially in dense areas [16]. Consequently, exploiting the spatial diversity, that is, several base stations covering the same device, is a way to lower the scheduling and queuing delays by dynamically selecting the base station with the smallest instantaneous load or replicating the packet on several base stations.

1.3 Thesis Outline and document structure

This manuscript is organized as follows. We first briefly present the evolution of mobile networks from a distributed to a centralized architecture and eventually to the advent of network slicing as a 5G systems enabler in

Chapter 2. We present the business relationships between the various actors in the 5G ecosystem and service level agreements between them.

In chapter 3, we identify the role of each player in network slice management and the entities that are responsible for accommodating the traffic and resources for the slices. Then, we describe the options for slice management function placement before delving into the challenges of URLLC services and an array of enablers for these cutting-edge use cases. Parts of the content presented in chapters 2 and 3 were published in the 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) conference proceedings [17].

In chapter 4, we study the performance of packet scheduling schemes for URLLC services. In order to ensure a low queuing time, we exploit the redundant coverage in many 5G RAN scenarios, where two RATs with the same service time are integrated. We consider three packet scheduling and redundancy schemes, namely Join-the-Shortest-Queue (JSQ), systematic Redundancy (RED), and redundancy with Cancellation upon completion (CAN). Based on queuing theory analysis, we develop expressions for reliability, defined as the probability that the packet is transmitted before some given target delay. We show that RED performs well at low load, while JSQ is better when the load increases; CAN outperforms all other schemes. We then show how the obtained results can be used to dimension the needed 5G radio resources and discuss the trade-off between performance and implementation complexity. The content of chapter 4 was presented at the International Teletraffic Congress (ITC 32) and published in its proceedings [18].

In chapter 5, we exploit redundant coverage of users to examine the impact of multi-connectivity in the case where base stations have different service capacities. Similarly to chapter 4, we study the performance of packet scheduling schemes for URLLC services. Thus, we review four packet scheduling and redundancy schemes, namely the JSQ, RED, CAN as mentioned before, and Shortest Expected Delay (SED).

Chapter 6 of this dissertation assesses the impact of architectural choices discussed in chapter 3 on the quality of service of different services, focusing on ultra-reliable low-latency communication applications. Our numerical experiments show that the placement of slice management functions plays a crucial role in choosing the radio resource allocation scheme that best fits

URLLC slices in both the homogenous and the heterogeneous cases. We also study the coexistence of URLLC and eMBB traffic to quantify the impact of serving the first on the latter. The content presented in this chapter was published in MDPI computers in the special issue "Selected Papers from 16th Wireless and Mobile Computing, Networking And Communications (WiMob 2020)" [19].

Chapter 7 summarizes the ideas presented in this dissertation and proposes possible research ideas to explore.

CHAPTER 2

EVOLUTION OF MOBILE NETWORKS TOWARDS NETWORK SLICING

Contents

2.1 From distributed to centralized radio access networks 8
2.2 From 4G C-RAN to 5G 10
2.3 From monolithic to slicing-empowered systems 12
2.4 5G ecosystem: business relationships and service level agreements 14
2.4.1 Business relationships between actors 14
2.4.2 Service Level Agreements 15
2.5 Summary 17

As mentioned before, 5G systems go beyond the one-size-fits-all architecture that characterized legacy networks to a more versatile network that can provide seamless data throughput, ultra-reliable low-latency applications, and offer support for a variety of IoT devices. Accordingly, this chapter summarises the evolution of the architectural design of RAN and how slicing-empowered networks are vital to meeting the novel applications' requirements.

According to [23], the number of Internet users is predicted to expand from 3.9 billion in 2018 to 5.3 billion by 2023, and the number of devices connected to IP networks is expected to achieve three times the global population by the same year. Furthermore, mobile devices and connections will grow to 12.3

billion by 2022, generating 77 Exabytes (EB) of mobile traffic [23]. Note that 1 EB = 10^{18} bytes, to illustrate, to have recorded 1 EB of data, you would have had started a video call 237823 years ago, which is comparable to the emergence date of homo sapiens on the planet [24].

This explosion in mobile traffic is due to the increasing numbers of customers using data-centric devices. Consequently, mobile network operators try to ensure the rising demands by exploring new solutions. Altering the network architecture is among the possible solutions. Yet, these adjustments are constrained by CAPEX and OPEX costs. Radio Access Network (RAN) is the part that resides between the user and the core network. It guarantees the user's access to the different services assured by the mobile network. It's composed of a Radio Frequency (RF) antenna located at the top of Base Station (BS), a Remote Radio Head (RRH), and a Baseband Unit (BBU) composed of dedicated equipment that may be placed at the bottom of the BS, or elsewhere serving a limited surface called a cell, as illustrated in Figure 2.1. The RF antenna, the RRH, and BBU connect through electrical cables, thus inducing degradation in signal transmission [25, 26].

2.1 From distributed to centralized radio access networks

RAN has evolved throughout the generations from traditional BS kept as one entity to Distributed RAN (D-RAN) and Centralized RAN (C-RAN). The latter is based on the idea of separating the analog processing parts, that is, the RRH also called the Radio Unit (RU) from their digital counterparts, that is the BBU or the Digital Unit (DU)[27]. Different types of connections can operate between the RRH and BBU, like fiber optics, coaxial cables, and even wireless connections [28].

The D-RAN architecture was initially implemented in the fourth generations' BS called Evolved Node B (eNodeB). This architecture introduced the separation between the DU and the RU. The DU is connected to the 4G core called the Evolved Packet Core (EPC) through the aggregation network. For example, in a downlink scenario, the BBU performs baseband signal processing converting the packets received from the EPC into a baseband signal.

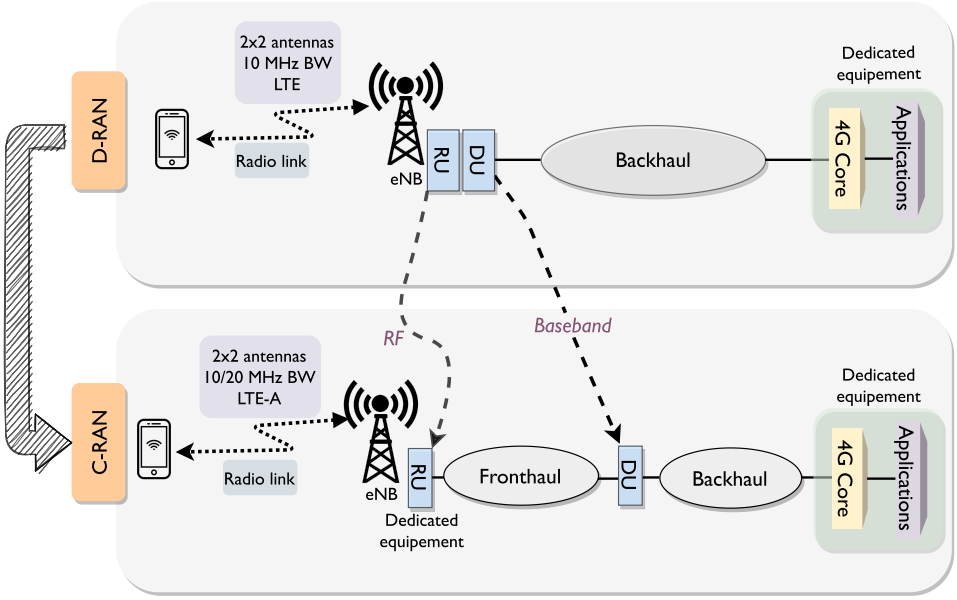


Figure 2.1: Evolution of cellular networks from distributed RAN to centralized RAN (adapted from [1]).

The baseband signal is then transmitted to the RU that carries out the digital processing (see Figure 2.1). The data transmission between the DU and RU is carried on using a radio protocol called the Common Public Radio Interface (CPRI). D-RAN represents several advantages. For instance, the RU can be positioned near the antenna on towers and rooftops, resulting in significant cuts in operational costs. However, RUs are assigned to DUs via dedicated links, which means they can not share the DU's processing capabilities resulting in underutilization of resources. Consequently, C-RAN has arisen to alleviate the inconveniences presented by the D-RAN architecture by putting together the DUs, therefore optimizing resource utilisations and energy efficiency.

C-RAN has been proposed to tackle the resources underutilization at the DU, introduced by D-RAN. While the RU remains at the BS, the DUs are gathered in one location called a central office using cloud infrastructures. The centralized DUs called DU pool or BBU pool allow a dynamic utilization of resources [29]. It performs the baseband processing of signals coming from different BSs. The link between the BBU and the RRH called the fronthaul

is exclusively a high-bandwidth and low-latency optical link in the wireline case [30, 31]. Moreover, the backhaul links the DU Pool with the mobile core network through the aggregation network. One of the main inconveniences presented by this architecture is the required CPRI rate. In particular, the required CPRI rate is 2.46 Gbps for a 20 MHz LTE system with two transmit antennas [32] and the overall CPRI rate increases linearly with the number of transmit antennas and bandwidth [33]. As a consequence, compression methods of the CPRI load were considered to relieve the cumbersome fronthaul requirements for infrastructure providers and telecom operators [34].

2.2 From 4G C-RAN to 5G

While C-RAN architecture presents numerous advantages, it strains the fronthaul link aggravating the requirement needed to transport the traffic. In [2], 3GPP defined functional splits of the protocol stack to alleviate this issue, prompting a new flexible architecture for the 5G RAN. The BBU functionalities are divided into a CU and Distributed Unit (DU) (to distinguish from the digital unit (DU) also known as the BBU, which we use in the following to avoid any confusion), where the latter is placed near the RU. In Figure 2.3, the combination of the RU and the DU is called the Access Unit (AU).

The separation between the RU, DU, and CU varies based on the chosen functional split subject to use cases and architecture proposals such as [35, 36], as illustrated in Figure 2.2. For example, option 7 marks the division between the DU and the RU, consisting of dividing Physical Layer (PHY) processing functionalities in the BBU pool and placing them near the RU. Similarly, option 2 representing the split between the Packet Data Convergence Protocol (PDCP) and Radio Link Control (RLC) characterize the CU/DU separation.

Besides, alleviating the bottleneck at the fronthaul link is essential to effective RAN deployment. The enhanced Common Public Radio Interface (eCPRI) standard has emerged as a successor to CPRI defined for C-RAN configuration. This protocol makes more efficient bandwidth use than its precursor [37]. It can also be framed within Ethernet since it is packet-based. Depending on the functional split, the fronthaul network can use Ethernet connectivity instead of relying on fiber optics links, bringing con-

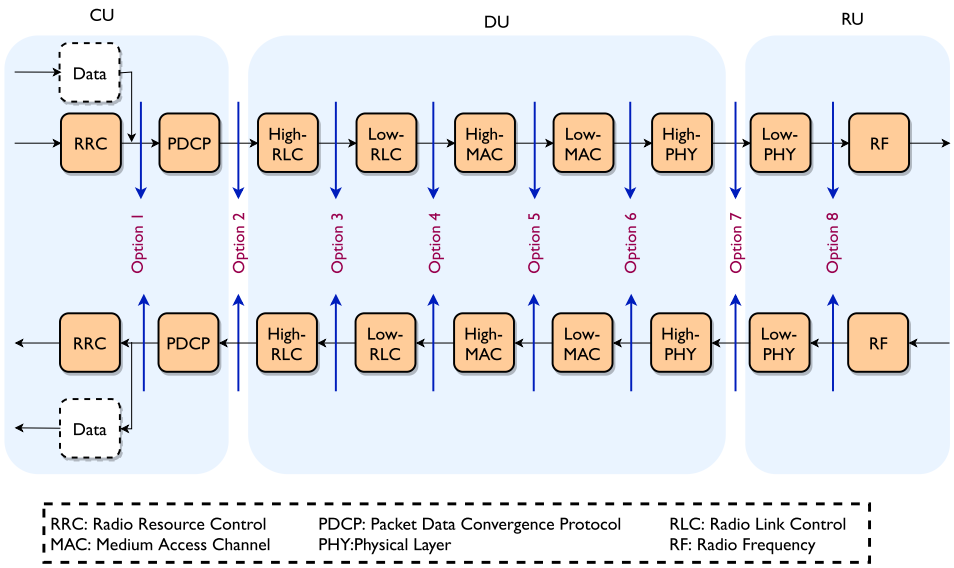


Figure 2.2: Functional split options between centralized, distributed and radio units, as defined by 3GPP [2].

siderable advantages. eCPRI is also an open interface, enabling scalability to network operators, capable of exploiting a variety of vendors' equipment.

5G networks promise a myriad of enhancements compared to previous generations, depending on the families of services introduced in chapter 1. Namely, 5G networks deliver high peak and user experience data rates up to 20 Gbps and 100 Mbps [38], respectively; enabled by the advent of massive Massive multiple-input multiple-output (MIMO), use of large bands of millimeter-wave spectra [39], and high spectral efficiency [40]. Further, 5G provides low latencies, down 10 times than 4G [41] and network energy-efficient consumption [42].

Accordingly, NFV backs 5G systems' promise since it allows the implementation, via virtualization, of Network Functions (NFs) as software rather than installing dedicated and proprietary equipment [43]. Hence, it enables a reduction in deployment costs as well as coexistence with legacy networks due to flexibility in deployment scenarios. In contrast, the main objective of SDN is to make networks more flexible and agile. The idea is to design, assemble, and operate networks that separate the network's control and user planes

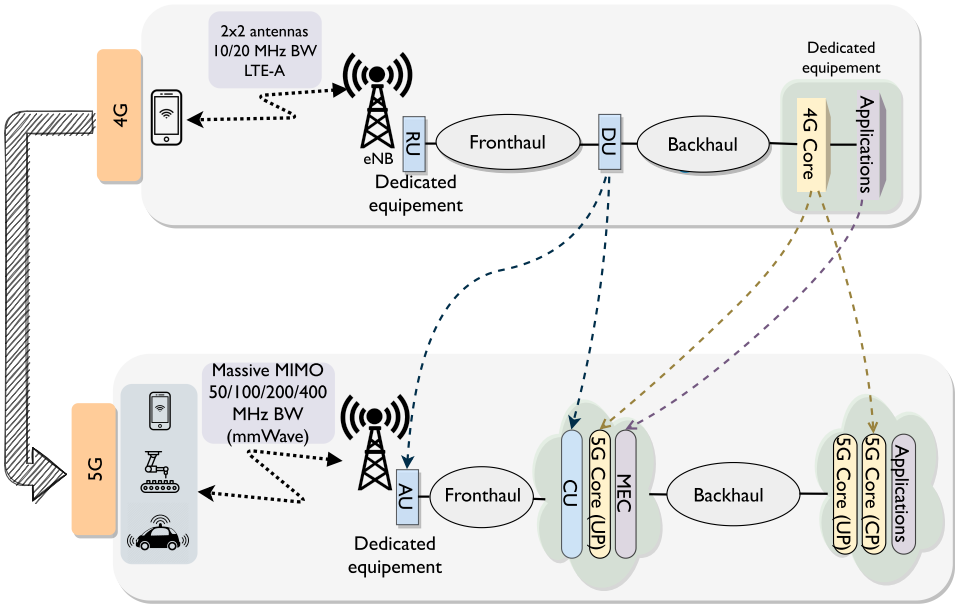


Figure 2.3: Evolution of cellular networks from centralized RAN to NFV/ SDN empowered 5G networks (adapted from [1]).

(also called forwarding or data planes), thus allowing the network control to become directly programmable and automated via a controller [44].

In short, NFV allows 5G RAN and core network implementation as software on commercial servers, whereas SDN permits network connectivity among virtual machines. Another advantage of NFV and SDN networks is disseminating cloud computing services to the network edge, called Mobile Edge Computing (MEC). These application servers are located at the network edge near end-users to supply low-latency services [45], as illustrated in Figure 2.3.

2.3 From monolithic to slicing-empowered systems

The combination of NFV and SDN brings several advantages for network operators, such as network programmability, energy efficiency, and delay reduction. For example, to improve C-RANs, NFV has been used to virtu-

alize the RAN architecture (vRAN) [46]. Authors in [47] tackle the vRAN design problem, proposing a common vRAN/MEC analytical framework, called FluidRAN. It minimizes RAN costs by jointly selecting the splits and RUs-CUs routing paths.

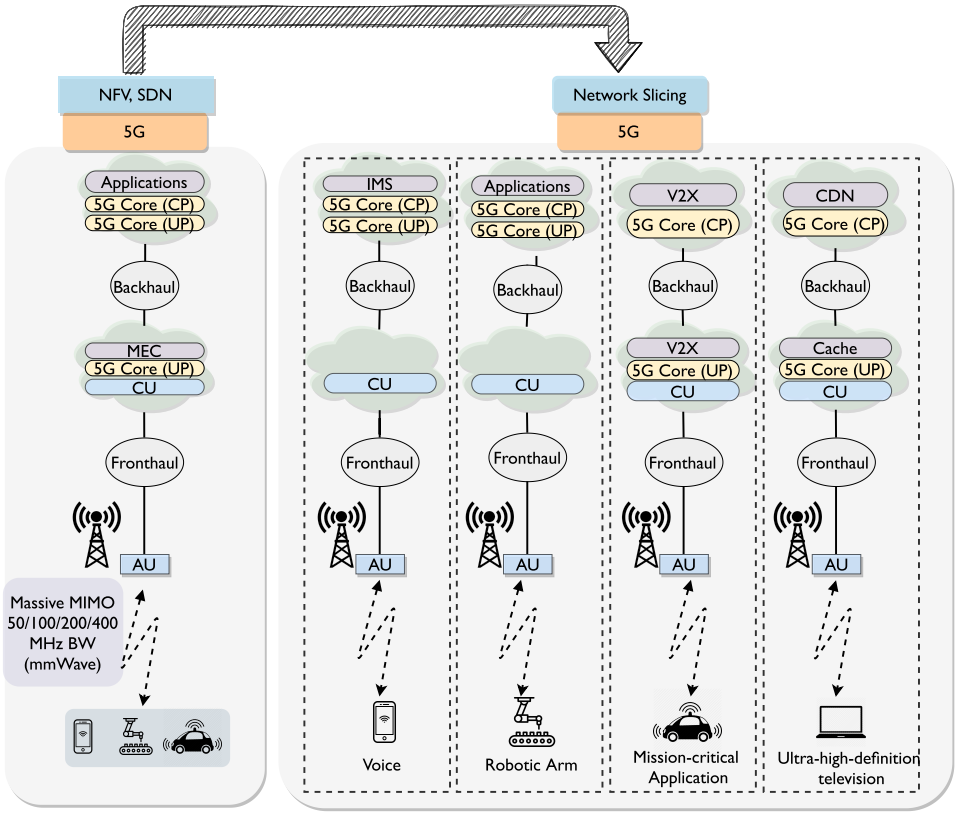


Figure 2.4: Evolution of cellular networks from NFV/SDN to network slicing based network (adapted from [1]).

[48] profits from the flexibility of virtualized RAN functions to introduce the RAN-as-a-Service (RANaaS) concept where RAN functions are centralized on a cloud computing platform, whereas authors in [49] demonstrate an example of RANaaS deployment using OpenAirInterface [50] and OpenStack.

Although the vision and targets of 5G have been extensively discussed, some of the research questions regarding the 5G networks infrastructure and ecosystem, enabling technologies, and application scenarios are yet to

be answered. As mentioned before, the 5G systems will support various new use cases from vertical industries, which inflict a much more comprehensive range of performance and cost requirements than traditional mobile networks. The current networks based on conventional "one-size-fits-all" design are not adaptable and scalable sufficiently to manage these diverse requirements in terms of performances, availability, security, and cost.

Network slicing has been proposed by academia and industry as a key enabler to support customized 5G network services on-demand, to handle a plethora of vertical-specific services alongside enhanced mobile broadband service over a shared physical network infrastructure [13]. This concept has emerged due to the recent advancement in NFV and SDN technologies. By slicing a physical network into several logical networks, each can provide tailored services for a characteristic application scenario [11]. As illustrated in Figure 2.4, 5G network slices represented by logically isolated and self-contained networks are flexible enough and highly customizable to deliver diverse business-driven use cases simultaneously over the same network infrastructure. Hence, it is critical to break the existing large monolithic network functions from legacy hardware into numerous software-based smaller functionality blocks with varying granularity; to achieve expected network services efficiently. Such cloud-native functionalities can be chained in flexible ways to form different network slices supporting 5G requirements.

2.4 5G ecosystem: business relationships and service level agreements

5G introduces new actors and business opportunities compared to previous generations [51]. Therefore, it is crucial to understand this ecosystem, identify players, and comprehend their roles in RAN slicing and resource allocation.

2.4.1 Business relationships between actors

The telecommunication industry is set to become a pedestal for a myriad of economic sectors. The Mobile Service Provider (MSP) plays a pivotal role in

this ecosystem, working as a mediator between the Infrastructure Provider (InP) and the tenants. The InP owns and manages the underlying resources, virtualized to build customizable E2E logical networks. The tenants can either be vertical actors, Mobile Virtual Network Operators (MVNO), or Over-The-Top (OTT) service providers.

The MSP leases resources (radio, processing, storage, and networking) from one or multiple InPs. Nevertheless, it may deploy its proper infrastructure and hence play also the role of InP, as illustrated in Figure 2.5. Additionally, the MSP offers and manages network services like eMBB slices to end-users. It also provides the necessary resources to carry on multi-tenancy scenarios, defined as the ability to supply various services to multiple tenants and to pool resources from several InPs at the same time [52]. Following the tenant's size and expertise, the latter may have control of the deployed NFs and can tailor the slice through its management and network orchestration (MANO) functional layer. Some verticals may even play the role of MSPs and manage their slices, as shown in Figure 2.5 [13].

In [12], the authors introduce the Network slicing as a service (NSaaS) paradigm. This concept allows operators to offer customizable E2E networks as a service. Accordingly, the MSP should enable access to a catalog of network slice templates stored in a repository to its clients (whether they are end-users or tenants). The network slice template emphasizes the structure of the network slice instance, specifying the necessary virtual and physical NFs (VNFs and PNFs) for deploying an E2E network slice, based on the technical requirements and constraints of the desired service.

In summary, the MSP leases chunks of the acquired resources from the InPs to the end-users and tenants. In the same way, the tenants provide the necessary resources to grant access to end-users through slices.

2.4.2 Service Level Agreements

Figure 2.5 represents the business relationships between actors and illustrates the complexity of resource management when it comes to ensuring QoS for slices. In other words, when SLAs are set up between the different actors, who is responsible for ensuring that their terms are respected?

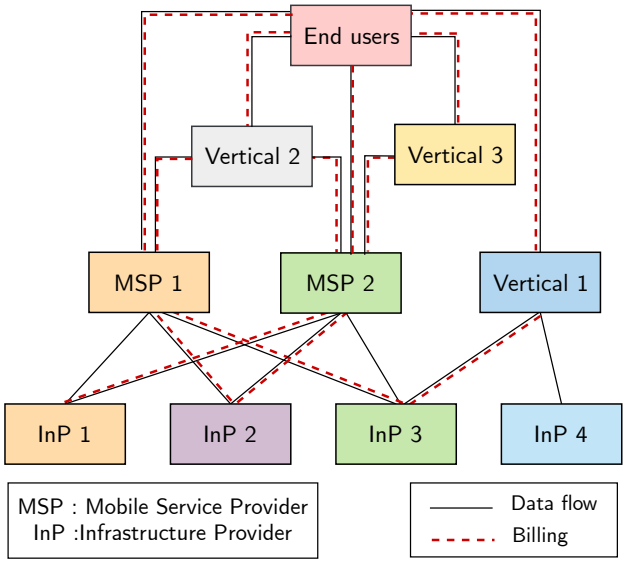


Figure 2.5: Business relationships in the 5G RAN. Solid lines correspond to data flow while dashed lines correspond to money flow. MSPs have billing relationships with InPs except when the InP is owned by the MSP.

One key point is that the InP cannot accept all resource demands from MSPs, as its resources are limited. In contrast, even if MSPs can lease resources from different InPs, they cannot admit every slice request from tenants, mainly because some slices are resource-hungry and thus burdensome to maintain. Keeping in mind that the InPs' objective is to maximize their revenue and resource utilization while the MSPs aim to maximize their revenues from tenants and minimize their resource leasing costs, guaranteeing QoS is not an easy task. Thereby, it is crucial to estimate the resources necessary to meet the slice SLAs. Note that the negotiated SLA is valid for the entire life-cycle of the network slice. It stipulates the customer-centered requirements stated in the form of Key Performance Indicators (KPIs) like capacity, reliability, availability, latency, and coverage area. Given the rising number of customers and applications, building customizable and programmable network slices requires an SLA management framework that automatically generates SLA templates and maps the high-level service-oriented requirement to a low-level technical description.

Many tenants do not have sufficient expertise in the telecommunications

field. Consequently, the MSP will have to guarantee the performance of the network slice agreed in the SLA on behalf of the tenant. In this case, the tenant has neither control nor visibility over the resources but receives performance reports to make sure that the service requirements specified in the SLA are respected. The tenant can demand some guarantees from the service provider in case the latter fails to deliver the high-level metrics agreed upon in the SLA. For instance, the tenant can apply a maximum penalty when the desired service is delay-sensitive as in URLLC applications. The clients can also negotiate the terms for putting an end to the negotiated SLA before the end of the life cycle.

Concerning the relationship between the MSP and the InP, it cannot incorporate a complete SLA as the InP does not have visibility on the resources allocated to the slice on several InPs. However, the MSP and the InP may establish a contract that stipulates the resource cost and a target acceptance ratio of resource allocation demands. The design of such contracts between MSPs and InPs, knowing the SLAs between tenants and MSPs, is an important research topic in the slicing context.

Moreover, multi-tenancy enabled networks are one of the important features to deliver 5G. Consequently, security and isolation are considered among the challenges to delivering E2E network slices rise. [53] presents a generic network slicing framework that includes management and operations-related mechanisms to address the multi-tenancy issue and cope with slice management scalability. These mechanisms are embedded in each slice using NFV MANO for slice orchestration and support multi-domain slicing.

2.5 Summary

In this chapter, we reviewed the evolution of RAN architecture from distributed to centralized networks, then to SDN/NFV networks, which has led to the advent of network slicing, which is key to delivering the novel 5G services promised by new actors. We also recapitulated the 5G ecosystem's players and their impact on resource management and summarized the possible business relationships between these actors and the stipulated SLAs between them. Accordingly, a natural inquiry would be to pinpoint the players' responsibilities in managing resources for network slices.

CHAPTER 3

NETWORK SLICING ARCHITECTURE

Contents

3.1 Slice management functions description 18
3.1.1 Resource allocation from MSP perspective 20
3.1.2 Resource allocation from tenant perspective 21
3.1.3 Resource allocation from InP perspective 21
3.2 Impact of Placement of intelligent entities . . . 22
3.2.1 Intelligence placed at the level of a shared RAN NSSMF 23
3.2.2 Intelligence placed at the NSMF level 24
3.3 Summary 25

Before describing the options for slice management function placement, we aim in this section at identifying the role of each player in the slice management and the entities that are responsible for managing traffic and resources for the slices.

3.1 Slice management functions description

The MSP has to create and simultaneously maintain many Network Slice Instances (NSIs). An NSI is composed of Core Network (CN) and RAN Network Slice Subnet Instances (NSSIs), arranged to provide necessary resources and functionalities and thus deliver the tenants' services. Each NSSI

encompasses PNFs and VNFs that are either dedicated or shared among different slices. According to the solution advocated by 3GPP and illustrated in Figure 3.1, the tenant's management function called the Communication Service Management Function (CSMF) forwards the service requirements to the Network Slice Management Function (NSMF). Then, the NSMF translates the E2E high-level performance requirements desired by the tenant to CN and RAN low-level requirements managed by the Network Slice Subnet Management Function (NSSMF). Subsequently, the RAN NSSMF converts the low-level requirements into Radio Resource Management (RRM) specific requirements and sets the resource allocation policy at the MAC scheduler of the Base Station (BS), whereas the CN NSSMF deploys and maps the service-oriented VNFs. Both the RAN and CN NSSMFs send periodic performance reports to the NSMF so that it can verify that the service requirements are respected. For example, if the RAN NSSMF violates the latency requirement of a network slice, the NSMF can adjust the scheduling policy by reserving more resource blocks. It can also alter the admission control procedure by rejecting any other network slice requests as long as the served slices SLAs are not respected. Table 3.1 summarizes these entities, their owners among the actors defined in chapter 2, and their roles in the slice resource allocation.

Function	**Location**	**Functionality**	**Owner**	**Autonomy**
UE scheduler	UE	Dispatches UE traffic to access points	Vertical	Applies policies specified by the vertical
BS scheduler	Base station	Allocates time/frequency resources to UEs	InP	Applies policies specified by the InP
NSSMF	RAN (e.g. Cloud RAN)	Orchestrates RAN resource allocation to slices	InP	Defines policies for the InP BSs
NSMF	MSP management server	Defines traffic steering policies for the slice	MSP	Defines MSP policies
CSMF	Tenant management entity (e.g. application Server)	Updates slice requirements and SLAs	Tenant	Defines tenant policies and needs

Table 3.1: Entities involved in RAN resource allocation and their roles.

Note that the resource allocation task is particularly complicated in case several MSPs lease resources from multiple InPs. Indeed, the NSMF belongs to the MSP and has as objective to ensure that the tenant's SLA is respected. Nevertheless, there is a RAN NSSMF that belongs to each InP, which has control over the resources of this particular InP only, as illustrated in Figure 3.1. In the latter, we consider three slices belonging to three different tenants. For each slice, we deploy an NSSMF per InP RAN. The question here is how to design and implement resource management policies in such

a distributed architecture while taking into consideration tenant, MSP, and InP perspectives.

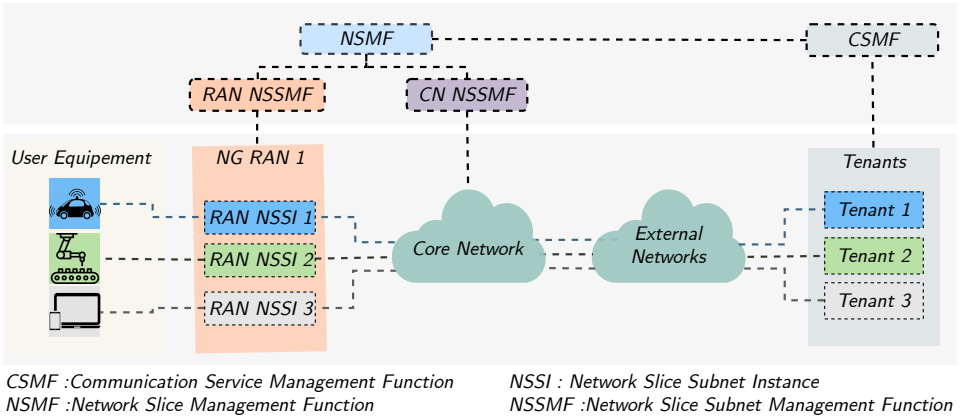


Figure 3.1: Overview of the infrastructure and management Layer in a network slicing scenario.

3.1.1 Resource allocation from MSP perspective

From the MSP perspective, the NSMF translates the tenant requirements into a traffic steering policy that determines to which InP(s) the packets of a specific User Equipment (UE) are to be forwarded. Such a policy may be generic, *e.g.* to privilege a particular InP when possible. Alternatively, it can be context-aware, which means examining the instantaneous load of the BS pertaining to an InP and its radio conditions with respect to the UE. For example, a potential policy is to connect a particular UE to a single InP, split its packets between several InPs, or even duplicate them to increase reliability. Specifically, the NSMF can, for example, decide that 70% of generated packets go through the main InP while the remaining 30% go through secondary InP during the validity time of the high-level policy.

If the MSP applies the decided policy without coordination with the InPs, an entity hosted in the UE capable of implementing the NSMF scheduling strategy is required. Otherwise, the traffic steering policy can be implemented as a shared NF among multiple slices on the InPs infrastructure or as dedicated NF with some cooperation between slices to meet the heterogeneous

optimization targets and for effective use of the radio resources [54].

3.1.2 Resource allocation from tenant perspective

From the tenant perspective, the CSMF determines dynamically the amount of resources that need to be allocated to the slice for continuously respecting the SLA, knowing the current traffic demand. In order for these requests to be accurate, the CSMF has to rely on the information originating from the application server and/or from the end-users. The time scale for these traffic reports has to be larger than the actual scheduler time scale, *i.e.* in the order of tens of milliseconds. In the specific case where the tenant is a "big" vertical that can deploy its own infrastructure (e.g., railway and highway management companies), it has the ability to bypass the MSP and acquire the resources directly from InPs, having thus the same behavior of MSPs, described previously.

3.1.3 Resource allocation from InP perspective

From the InP perspective, the NSSMF receives the resource allocation requests from the UEs belonging to different slices and applies some scheduling/admission control policies to them. The devised policies of the InP have to dynamically share the resources among the slices to raise the overall resource efficiency, especially that leasing fixed shares of resources will limit the multiplexing gains.

Note that, from an InP perspective, [55] introduces the so-called 5G network slice broker, hosted in the NSSMF of the InP, that gathers global network load measurements and configures the RAN scheduler policies based on the negotiated SLA and the size of the network slice. Moreover, the openness of the mobile network may lead to an adversarial behavior of MSPs consisting of maximizing the acquired share of resources. In order to deal with this issue, a 'share-constrained proportional allocation' mechanism is exploited in [56], and the share obtained by each tenant is determined by the equilibrium point of a network slicing game. In the same context, the authors in [57] investigate resource allocation mechanisms between tenants using game theory tools to model the non-cooperative behavior of slices. However, these works are

limited to multiple tenants sharing a single InP infrastructure.

3.2 Impact of Placement of intelligent entities on Radio Resource Allocation for slices

We now study the placement of entities in charge of resource allocation in light of the above description of the slice management functions. We consider, for illustration, the case of a smart factory where several BSs (5G NR and/or legacy) are deployed to establish a redundant coverage, essential for ensuring URLLC QoS, as illustrated in Figure 3.2. Note that, 5G NR and 4G BSs can natively cooperate via a common core network, whereas [58] prescribes the Non-3GPP Interworking Function (N3IWF) for combining accesses using proprietary or WiFi technology. The tenant may own and manage some small cells deployed within the factory, while the InP manages BSs, operating in the sub 2 GHz spectrum for ensuring full coverage.

While some UEs will be covered by the macro cells only, it is envisioned that most locations will be covered by at least two overlapping cells, providing flexibility in resource allocation and redundancy for ensuring reliability. We hereafter display three potential resource allocation schemes exploiting these advantages.

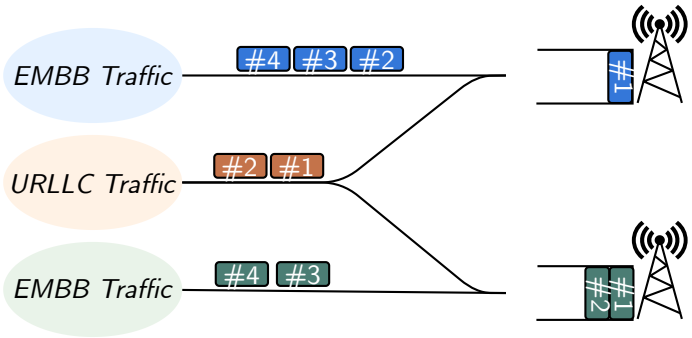


Figure 3.2: Traffic steering in industry 4.0 use case. Two BSs in the neighborhood of URLLC user equipment.

3.2.1 Intelligence placed at the level of a shared RAN NSSMF

In this case, the traffics of all RAN slices and the radio resources of all BSs are managed via a shared RAN NSSMF with a single compound MAC scheduler. The latter has access to real-time information concerning the time-frequency matrix of each BS, thus allowing grant-based resource allocation. This case is enabled when the resources of all BSs are centrally managed within a common Cloud-RAN linked to the BSs by a high capacity fronthaul, as illustrated in Figure 3.3.

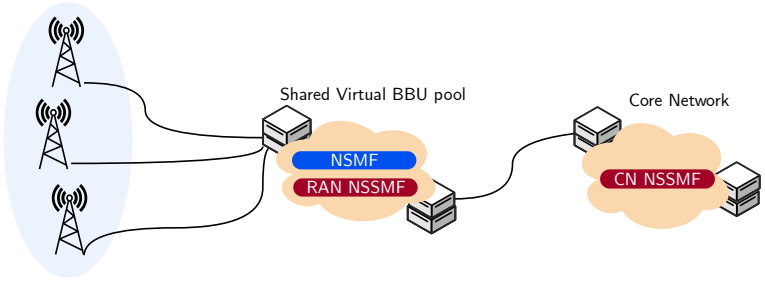


Figure 3.3: Distribution of management functions in a factory scenario with the intelligence placed at the level of a shared RAN NSSMF.

A dynamic strategy can thus be applied by the NSSMF to URLLC traffic, which consists of sending packets to the BS with the lowest instantaneous load in order to minimize latency. As of eMBB traffic, it is served by one of the two BSs independent of the instantaneous load, *i.e.* each BS has its own eMBB traffic to serve and manages the URLLC traffic jointly with the other BS.

This strategy can be applied in both uplink and downlink; it is straightforward in the downlink where the application server sends the URLLC packets to the NSSMF that directs them to the adequate BS for transmission. As of the uplink, the UE sends a scheduling request to NSSMF that issues a scheduling grant on one of the BSs. Consequently, the uplink case is more challenging as this control process may introduce latency between the moment the loads are observed by the NSSMF and the moment the scheduling grant is issued for the URLLC user.

3.2.2 Intelligence placed at the NSMF level

When there is restricted coordination between the InPs, and between the MSP and the InPs, as in the case where each BS has its own Baseband Units (BBUs), loosely linked to other BBUs, performing intelligent steering of each packet based on the instantaneous load of each cell is difficult to achieve. This is illustrated in Figure 3.4. In this case, a long-term policy (based on a time granularity of tens of seconds) is to be applied, managed by the NSMF located somewhere at the level of the core network. For this policy to be effective, the UEs (in the uplink) and the application server (in the downlink) have to apply the policy provided by the NSMF on a packet basis, but without further information on the instantaneous load of each cell.

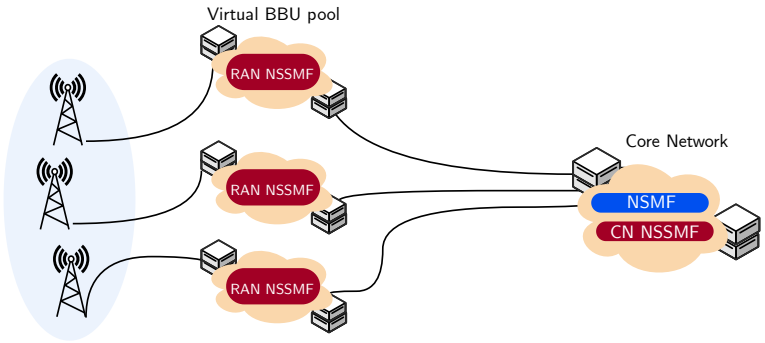


Figure 3.4: Distribution of management functions in a factory scenario with the intelligence placed at the NSMF level.

When the NSMF takes the decision about the destination of the packet, the remainder of the scheduling process is performed classically, and the RAN NSSMF does not need to know about the slice policy. We consider hereafter two feasible policies for URLLC:

1. Long-term traffic steering with no redundancy: It entails the division of the URLLC traffic proportionally, based on the base stations' average capacities as estimated by the NSMF, or as provided to the MSP by the RAN NSSMF of each InP.

2. Long-term traffic steering with redundancy: In the absence of any information about the capacities of the different BSs, and in order

to ensure reliability, redundancy is a costly yet simple strategy. This implicates sending systematically the arriving URLLC packets to both BSs. While packet redundancy can achieve high reliability as it enables the experience of minimum queuing latency between the BSs, it leads to the under-utilization of radio resources. The NSMF broadcasts the policy to the URLLC user equipment during the slice instantiation.

3.3 Summary

In this chapter, we presented the entities implicated in RAN resource allocation and their respective roles while offering a thorough description of slice management functions. Then, we explored the network slicing architecture's impact on URLLC performance in the multi-connectivity case. Depending on chapter 2 ecosystem description, we studied various options for the placement of the management entities involved in resource allocation and traffic steering decisions while focusing on the challenging use case of URLLC traffic. In particular, two architectural options have been studied; the first one is with loose coupling, where the scheduling policy is determined within the NSMF by the vertical. The second one uses tight coupling where the scheduling decision is taken at the NSSMF level by the mobile network operator who owns the different base stations.

CHAPTER 4

MULTI-CONNECTIVITY FOR URLLC SLICE IN HOMOGENEOUS SETTINGS

Contents

4.1 System Model 27
4.2 Main results 29
4.2.1 Occupancy distributions 29
4.2.2 Decay Rates 32
4.2.3 Outage Probability 38
4.2.4 Average Sojourn Time 43
4.3 Resource dimensioning 44
4.4 Summary 46

Having studied in the previous chapter the impact of architectural choices on the URLLC slice performance, we explore in this chapter in more details the impact of specific scheduling policies on the performance. As before, we exploit the redundant coverage in many 5G RAN scenarios, where two frequency layers or RAT are integrated. We consider three packet scheduling and redundancy schemes, namely Join-the-Shortest-Queue (JSQ), systematic Redundancy (RED), and Redundancy with Cancellation (CAN). On the basis of queuing theory results, we develop expressions for the reliability, defined as the probability that the packet is transmitted before some given target delay.

4.1 System Model

We consider the downlink of a wireless system with a set of URLLC users located within an area served by two RATs. The RAT's may belong to different InPs, but they are able to serve dynamically packets belonging to URLLC users. A possible architecture that allows this dynamic service of packets is the one described in Figure 4.1, where an entity connected to the two BSs is responsible for dispatching/duplicating packets. This dynamic packet scheduling is performed on the basis of the instantaneous system state, following one of the policies outlined below. Note that, if 5G slicing is implemented, the RAN slice manager (also defined in 3GPP as the NSSMF [59]) may be responsible for this dynamic management as follows: it receives the application packets from the application server belonging to the vertical (slice owner) and sends them to the schedulers of the BSs. This decision is based on periodical updates received from the schedulers of the different BSs about their load status.

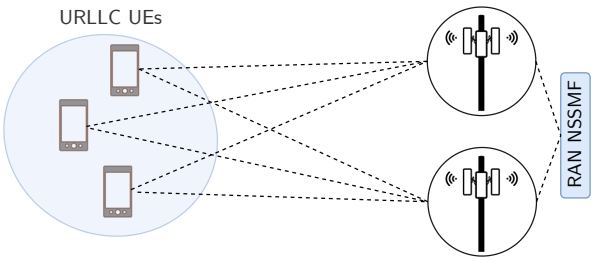


Figure 4.1: Two BSs in the neighborhood of URLLC user equipment.

When a packet belonging to a URLLC device arrives at the scheduler, three different policies can be applied:

- **Join-the-Shortest Queue discipline:** the first scheme consists of sending the incoming URLLC packet to the queue with the least number of waiting packets. If both BS's are empty or have the same number of waiting packets, packets are equally likely to join either BS;
- **Redundancy discipline:** each incoming packet is independently duplicated in both queues. This scheme does not require any prior knowledge of the radio access channel; thus there is no need for extensive

control plane information;

- **Redundancy with Cancellation discipline:** as in the previous case, we send the incoming packet to both BS's. This scheme entails the elimination of the remaining copy, provided that one of the copies has been fully served.

We denote the aforementioned schemes by **JSQ**, **RED** and **CAN**, respectively.

We model the network architecture by two parallel queues fed by a Poisson process of URLLC packets with mean arrival rate λ, the size of packets being denoted by W (bytes). Motivated by the flexibility of the 5G NR air interface, we consider a First Come First Serve (FCFS) discipline for each queue. This means that the BS adapts its mini-slot dynamically so that one URLLC packet is served by the BS during one mini-slot [1]. Service times of packets at either queue are assumed to be mutually independent and exponentially distributed with identical rate α. While the packet size is the same on both queues as they correspond to replicas, the resources available for URLLC depend on the traffic load for other services on each BS, hence the independence assumption. In the following, we set

$$\varrho = \frac{\lambda}{2\alpha}. \tag{4.1}$$

Given these two $M/M/1$ queues coupled by either JSQ, RED or CAN discipline, we denote by M (resp. N) the number of packets in the first (resp. the second) queue. The associated stationary distribution of the occupancy vector (M, N) is then defined by $\Pi_{m,n} = \mathbb{P}(M = m,\, N = n)$, $(m, n) \in \mathbb{N}^2$; the service rate α being identical at each queue, this distribution is symmetric, that is, $\Pi_{m,n} = \Pi_{n,m}$ for any pair (m, n). Following [60], [61] and [62], respectively, this stationary distribution is then shown to exist provided that

- for JSQ, $2\alpha > \lambda$, that is, $\varrho < 1$;
- for RED, $\alpha > \lambda$, that is, $\varrho < 1/2$;

[1] Note that, in cases where the amount of spectral resources is large, and the packet is small, several packets may be multiplexed in the frequency dimension in the mini-slot of smallest size (2 OFDMA symbols). Our assumption of a FCFS rule for each queue then gives an upper bound of the performance, assuming a maximal slot size flexibility.

- for CAN, $2\alpha > \lambda$, that is, $\varrho < 1$, identical to the stability condition for JSQ.

4.2 Main results

For each of the above allocation schemes, the performance indicator is the reliability metric $\mathbb{P}(T \leqslant t_0)$, where T is the sojourn time of a packet in the system and t_0 is the delay budget. This can be completely characterized by the distribution of T which, however, is difficult to obtain explicitly for both JSQ and RED. In fact, it is closely related to the occupancy distribution $(\Pi_{m,n})$ which is only accessible through an intricate expression of its generating function. To compare the respective performance of the three schemes, nevertheless, we first determine the exponential decay rate at infinity of the distribution of T, that is, the positive limit

$$-\lim_{t\to+\infty} \frac{\log \mathbb{P}(T > t)}{t} \tag{4.2}$$

which is shown to exist for each scheme. When the reliability constraint is very strict, as for URLLC, t_0 is large compared to the average sojourn time and the decay rate is a good proxy for the outage rate, as will be illustrated in the numerical examples. Furthermore, we show how the full distribution of T can be numerically calculated by means of contour integrals. (see Appendix A)

4.2.1 Occupancy distributions

For both JSQ and RED schemes, we here recall useful results from earlier references regarding the joint stationary distribution $(\Pi_{m,n})$, $(m,n) \in \mathbb{N}^2$, of the occupancy vector (M,N). As above, the arrival rate λ is here normalized to 1 with $\varrho = 1/2\alpha$.

4.2.1.1 JSQ scheme

Let G denote the generating function of the occupancy distribution $(\Pi_{m,n})$, defined by

$$G(x,y) = \sum_{m\geqslant 0, n\geqslant 0} \Pi_{m,n} x^m y^n \tag{4.3}$$

for $|x| \leqslant 1$ and $|y| \leqslant 1$. Basic functional properties of G can be summarized as follows:

• first, the symmetry of distribution $(p_{i,j})$ enables one to express G as

$$G(x,y) = F(xy,x) + F(xy,y) - F(xy,0) \tag{4.4}$$

for $|x| \leqslant 1$ and $|y| \leqslant 1$, where the auxiliary function F is defined by

$$F(x,y) = \sum_{0\leqslant m\leqslant n} \Pi_{m,n} x^m y^{n-m}.$$

Besides, the probability $G(1,0) = G(0,1)$ that either queue is empty is given by $F(0,1) = 1 - \varrho$ with $\varrho = 1/2\alpha$.

• secondly [60], function F is determined by

$$F(x,y) = \frac{J(x,y)}{K(x,y)} \tag{4.5}$$

in terms of $F(x,0)$ and $F(0,y)$, where

$$J(x,y) = x\left(x + \alpha - (1+2\alpha)\frac{y}{2} - \frac{y^2}{2}\right) F(x,0) \;+\; \alpha y(y-x)F(0,y)$$

and

$$K(x,y) = x(x+\alpha) - (1+2\alpha)xy + \alpha y^2.$$

In the derivation procedure for unknown terms $F(x,0)$ and $F(0,y)$ in the numerator $J(x,y)$ of (4.5), it proves essential to deal with a rational parametrization $\mathfrak{p} \mapsto (X(\mathfrak{p}), Y(\mathfrak{p}))$ of the conic with equation $K(x,y) = 0$. Defining the mapping $\mathfrak{p} \mapsto X(\mathfrak{p})$ by

$$X(\mathfrak{p}) = \frac{a}{4}\left(\mathfrak{p} + \frac{1}{\mathfrak{p}}\right) + \frac{a}{2} \tag{4.6}$$

with coefficient $a = 4\alpha^2/(1+4\alpha^2)$, the unknown term $F(x,0)$ in (4.5) is then shown to read

$$F(X(\mathfrak{p}),0) = \frac{\alpha}{\alpha+1}\cdot\frac{D(\mathfrak{p})}{D(\mathfrak{p}_0)} \tag{4.7}$$

with function D expressed in terms of infinite products as [60, Theorem 2]

$$D(\mathfrak{p}) = \frac{\displaystyle\prod_{n\geqslant 2}(1+k^n\mathfrak{p})\left(1+\frac{k^n}{\mathfrak{p}}\right)}{\displaystyle\prod_{n\geqslant 0}\left(1-\frac{k^n}{k'}\,\mathfrak{p}\right)\left(1-\frac{k^n}{k'\,\mathfrak{p}}\right)} \tag{4.8}$$

and constants $\mathfrak{p}_0$, k, k' given by

$$\begin{cases} \mathfrak{p}_0 = \dfrac{1+2\alpha^2+\sqrt{1+4\alpha^2}}{2\alpha^2}, \\[2ex] k = \dfrac{1+2\alpha-\sqrt{1+4\alpha^2}}{1+2\alpha+\sqrt{1+4\alpha^2}}, \\[2ex] k' = 1+8\alpha^2+4\alpha\sqrt{1+4\alpha^2} \end{cases} \tag{4.9}$$

(the other component $Y(\mathfrak{p})$ and the associated derivation of the term $F(0,y)$ are not presently needed and thus omitted);

• the generating function for either M or N reads

$$G(x,1) = \frac{x(x+\alpha)}{2\alpha(x-\alpha)}\,F(x,0) - \frac{\alpha}{x-\alpha}\,F(0,1), \tag{4.10}$$

and thus depends on probability $F(0,1) = 1-\varrho$ and on function $x \mapsto F(x,0)$ only. To derive $F(x,0)$ from the expression (4.7) of $F(X(\mathfrak{p}),0)$, the relevant inverse $\mathfrak{p} = \mathfrak{P}(x)$ to the quadratic equation $X(\mathfrak{p}) = x$ for any given x is

$$\mathfrak{P}(x) = \frac{2x}{a} - 1 + \frac{2}{a}\sqrt{x(x-a)} \tag{4.11}$$

with constant $a = 4\alpha^2/(1+4\alpha^2)$ and where the branch of the square root is chosen so that $\sqrt{z} > 0$ for $z > 0$;

• finally, using Little's law (with arrival rate fixed here to unity), the mean sojourn time $\mathbb{E}(T)$ can be derived from the expression of the mean total

number of customers [60, Section 4], namely

$$\begin{aligned} 1 \times \mathbb{E}(T) &= \mathbb{E}(M) + \mathbb{E}(N) = 2\,\mathbb{E}(M) \\ &= \frac{\alpha}{\alpha^2 - 1} + \frac{1 + 4\alpha^2}{\alpha^2(1-\alpha)} \cdot \frac{\mathfrak{p}_0^2}{\mathfrak{p}_0^2 - 1}\, S(\alpha) \end{aligned} \tag{4.12}$$

where $\alpha = 1/(2\varrho)$ and

$$S(\alpha) = \sum_{n \geqslant 2} \left[\frac{k^n}{1 + k^n \mathfrak{p}_0} - \frac{k^n}{\mathfrak{p}_0(k^n + \mathfrak{p}_0)} \right] + \sum_{n \geqslant 0} \left[\frac{k^n}{k' - k^n \mathfrak{p}_0} - \frac{k^n}{\mathfrak{p}_0(\mathfrak{p}_0 k' - k^n)} \right]$$

with constants $\mathfrak{p}_0$, k, k' specified above in (4.9).

Compared to the original version of the paper [60], some typos have been corrected, namely: the infinite product in the denominator of $D(\mathfrak{p})$ in expression (4.8) should display two minus signs (and not a + and a -), and the denominator in the expression (4.9) of constant $\mathfrak{p}_0$ should be $2\alpha^2$ and not 2α.

4.2.1.2 RED scheme

Now turn to the RED scheme. As shown in [61, 63], the generating function H of the stationary distribution $(\Pi_{m,n})$ is given by

$$H(x,y) = \alpha \cdot \frac{x(y-1)H(x,0) + y(x-1)H(0,y)}{(1+2\alpha)xy - \alpha(x+y) - x^2y^2} \tag{4.13}$$

with

$$H(x,0) = H(0,x) = \frac{(\alpha - 1)^{\frac{3}{2}}}{\alpha\sqrt{\alpha - x}}$$

for $|x| \leqslant 1$ and $|y| \leqslant 1$.

4.2.2 Decay Rates

We first assert the following.

Proposition 1. ***The decay rate of the distribution of sojourn time*** T ***is given by*** $\alpha\,\Theta^*$***, where***

$$\Theta^* = \begin{cases} S^* = 1 - \varrho^2 & \textit{for JSQ} \\ \Sigma^* = 2\left(1 - \dfrac{2\sqrt{\varrho}}{\sqrt{4+\varrho} - \sqrt{\varrho}}\right) & \textit{for RED} \\ \Xi^* = 2(1-\varrho) & \textit{for CAN} \end{cases} \tag{4.14}$$

in terms of parameter $\varrho = \lambda/2\alpha$***.***

Proof. We successively calculate the decay rate for the distribution of the sojourn time T for the JSQ, RED and CAN scheme, respectively. Unless otherwise mentioned, the arrival rate λ is here normalized to 1; definition (4.1) of parameter ϱ thus reduces to $\varrho = 1/2\alpha$.

4.2.2.1 Decay rate for JSQ

a) Let T denote the sojourn time of a tagged packet entering the system and by S its service time. Recall that each individual queue is ruled by the "First Come First Served" (FCFS) discipline. Given the occupancy vector (M, N), T is then given by

$$T = S_1 + ... + S_K + S \tag{4.15}$$

where K has the distribution of either random variable M or N, since the distribution of (M, N) is symmetric; all random variables S_1, ..., S_K, S are mutually independent and identically distributed, with exponential distribution with mean $1/\alpha$. After (4.15), the Laplace transform φ^* of T is the $(K{+}1)$-fold convolution with a random number K, where K has the distribution of either random variable M or N. We consequently have

$$\varphi^*(s) = \mathbb{E}\left[\left(\frac{\alpha}{\alpha+s}\right)^{M+1}\right] = \frac{\alpha}{\alpha+s} \cdot G\left(\frac{\alpha}{\alpha+s}, 1\right)$$

for $\mathrm{Re}(s) \geqslant 0$; applying the expression (4.4) for $G(x,1)$ in terms of F then provides

$$\varphi^*(s) = \frac{\alpha}{\alpha+s}\left[F\left(\frac{\alpha}{\alpha+s},\frac{\alpha}{\alpha+s}\right) + F\left(\frac{\alpha}{\alpha+s},1\right) - F\left(\frac{\alpha}{\alpha+s},0\right)\right] \tag{4.16}$$

for $\mathrm{Re}(s) \geqslant 0$.

b) It is known that if the Laplace transform φ^* has a simple pole at some point $-s^*$, $\mathrm{Re}(s^*) > 0$, and with no other such poles with less module, then the limit (4.2) exists and is precisely s^*. It is therefore sufficient to determine the simple pole of φ^* with smallest module.

Following (4.16), the possible poles of φ^* are those of either $F(x,x)$, $F(x,1)$ and $F(x,0)$ where we set $x = \alpha/(\alpha+s)$ for short. Now, using (4.5), we easily calculate

$$\left\lfloor \begin{array}{l} F(x,x) = \dfrac{x+2\alpha}{2\alpha}\,F(x,0), \\ \\ F(x,1) = \dfrac{xF(x,0) - \alpha F(0,1)}{x-\alpha}. \end{array} \right. \tag{4.17}$$

We then successively observe that

• the singularities of $F(x,x)$ either correspond to the singularities of either x (that is, $s = -\alpha$) or $F(x,0)$;

• the singularities of $F(x,1)$ either correspond to $x = \alpha$ or to the singularities of $F(x,0)$ again. In the former case, recall that the vanishing of the denominator $K(x,1) = (x-1)(x-\alpha)$ for $x = \alpha$ must correspond to the vanishing of the numerator $J(\alpha,1)$ for $F(\alpha,1)$ to be well-defined; the point $x = \alpha$ thus cannot be a singularity of $F(x,1)$. We thus conclude that the only possible singularities of $F(x,1)$ are those of $F(x,0)$ only.

By the latter discussion, the only singularities of $\varphi^*(s)$ are either $s = -\alpha$ or those of $F(x,0)$ with $x = \alpha/(\alpha+s)$. Turning to the singularities of $F(x,0)$, we deduce from formula (4.7) for $F(X(\mathfrak{p}),0)$ and the associated expression (4.8) of $D(\mathfrak{p})$ that $F(X(\mathfrak{p}),0)$ is infinite if and only if $D(\mathfrak{p})$ is. After (4.8), this corresponds to the family of real simple poles $\mathfrak{q}_r$, $r \in \mathbb{N}$, and $\mathfrak{q}'_r$, $r \in \mathbb{N}$,

with

$$\mathfrak{q}_r = \frac{k'}{k^r}, \qquad \mathfrak{q}'_r = \frac{1}{\mathfrak{q}_r} = \frac{k^r}{k'}. \tag{4.18}$$

Variable change $x = \alpha/(\alpha + s)$, definition (4.6) for $X(\mathfrak{p})$ and the property $X(\mathfrak{p}) = X(1/\mathfrak{p})$ for $\mathfrak{p} \neq 0$, then determine the real simple poles $-s_r$, $r \in \mathbb{N}$, of $\varphi^*(s)$ given by

$$-s_r = -\alpha + \frac{\alpha}{X(\mathfrak{q}_r)}, \qquad r \in \mathbb{N}, \tag{4.19}$$

where

$$X(\mathfrak{q}_r) = \frac{a}{4}\left(\mathfrak{q}_r + \frac{1}{\mathfrak{q}_r}\right) + \frac{a}{2} \tag{4.20}$$

after (4.6). Noting from (4.9) that $k' > 1$, we readily derive from (4.18) that $\mathfrak{q}_0 > 1$ and the fact that $0 < k < 1$ entails that the sequence $(\mathfrak{q}_r)_{r\in\mathbb{N}}$ is strictly increasing. As the function $p \in [1, +\infty[\mapsto p + 1/p$ is strictly increasing, the pole of φ^* with smallest module thus corresponds to the index $r = 0$, that is,

$$-s^* = -s_0 = -\alpha + \frac{\alpha}{X(\mathfrak{q}_0)} \tag{4.21}$$

with $-\alpha < -s^* < 0$ (the point $s = -\alpha$ is therefore not the singularity of $F(x, 0)$ with least module).

Replacing constant $\mathfrak{q}_0 = k'$ by its expression (4.9) in terms of α and using (4.20) for $r = 0$, the relation (4.21) for s^* reads

$$-s^* = -\alpha + \frac{1 + 4\alpha^2}{\alpha(2 + k' + 1/k')}$$

where $1/k' = 1 + 8\alpha^2 - 4\alpha\sqrt{1 + 4\alpha^2}$. Writing $\alpha = 1/2\varrho$, the latter expression for s^* reduces to $s^* = (1 - \varrho^2)/2\varrho$ after some elementary algebra. Restating physical units, the decay rate αS^* of sojourn time T for JSQ is deduced from equality $\lambda\, s^* = \alpha\, S^*$ with $\lambda = 2\alpha\varrho$, hence $S^* = 2\varrho\, s^* = 1 - \varrho^2$ as claimed in (4.14) ■

4.2.2.2 Decay rate for RED

To derive the decay rate Σ^* for RED, an analytic proof similar to that of Section 4.2.2.1 above can be performed. We here prefer to provide probabilistic

arguments for this derivation; the latter invokes, in particular, a comparison with another queuing system with batch arrivals.

a) Given the occupancy vector (M, N), the sojourn time T of a tagged packet in the RED scheme is given by the minimum

$$T = \min\left(S_1 + ... + S_M + S,\ S'_1 + ... + S'_N + S'\right) \tag{4.22}$$

where S (resp. S') denotes the service time of the duplicated packet in the first queue ♯1 (resp. in the second queue ♯2); all random variables S_1, ..., S_M, S and S'_1, ..., S'_N, S' are assumed to be mutually independent and identically distributed, with exponential distribution with mean $1/\alpha$.

b) Consider the arrival time τ of a tagged packet; at that time, the time backlog of queue ♯1 (resp. queue ♯2) is V (resp. V'). Once duplicated, the packet brings an additional finite backlog S (resp. S') to queue ♯1 (resp. queue ♯2). Conditioned on the event $(T > t)$ with large t, we have $V = O(t)$ and $V' = O(t)$ while $S/V = o(1)$ and $S'/V' = o(1)$. Applying (4.22), the sojourn time of the packet is thus of order

$$T = \min(V + S, V' + S') \sim V \sim V'. \tag{4.23}$$

Now, consider an $M^{[X]}/M/1$ FCFS queue fed by a Poisson process with rate 1, with batch arrivals of constant size $B = 2$ and service rate 2α. At the same arrival time τ and with an identical total number of customers $M+N$, the time backlog of this queue is

$$\mathscr{T} = \frac{V + V'}{2} + \frac{S + S'}{2}$$

where the factor 2 stems from the fact that the service rate has been doubled; in view of (4.23), we then have

$$\mathscr{T} \sim \frac{V + V'}{2} \sim T.$$

We thus conclude that probabilities $\mathbb{P}(T > t)$ and $\mathbb{P}(\mathscr{T} > t)$ are of the same order for large t.

c) The stability condition for the $M^{[X]}/M/1$ queue with batch arrivals of size $B = 2$ reads $\lambda\,\mathbb{E}(B) = 2 < 2\alpha$ [64, Sect.4.1], that is, $\alpha > 1$ as required;

its stationary occupancy distribution has the generating function Q given by

$$\begin{aligned} Q(z) &= \frac{2\alpha Q(0)(1-z)}{2\alpha(1-z) - z(1-\mathbb{E}(z^B))} \\ &= \frac{2(\alpha-1)}{2\alpha - z(1+z)} \end{aligned} \tag{4.24}$$

for $|z| \leqslant 1$. The tagged packet has the first position within the batch of size $B = 2$ with probability $p = 1/2$; the Laplace transform ψ^* of its sojourn time $\mathscr{T}$ is consequently given by

$$\psi^*(s) = 2 \times \frac{2\alpha}{s+2\alpha} \cdot Q\left(\frac{2\alpha}{s+2\alpha}\right), \quad \mathrm{Re}(s) \geqslant 0, \tag{4.25}$$

(where the multiplying factor 2 comes from dividing by the conditional probability $p = 1/2$). Now, using (4.24), it is readily verified that the pole of smallest module of Q equals $z_+^* = (\sqrt{1+8\alpha} - 1)/2 > 1$; following (4.25), the pole of Laplace transform ψ^* with smallest module is thus given by $s = -\sigma^* = -2\alpha + 2\alpha/z_+^*$, that is,

$$\sigma^* = \frac{1}{\varrho}\left(1 - \frac{2\sqrt{\varrho}}{\sqrt{4+\varrho} - \sqrt{\varrho}}\right)$$

in terms of $\varrho = 1/(2\alpha) < 1/2$. Restating time units, the decay rate $\alpha\Sigma^*$ of T for RED follows from equality $\lambda\,\sigma^* = \alpha\,\Sigma^*$ with $\lambda = 2\alpha\varrho$, which provides expression (4.14) for Σ^* ■

4.2.2.3 Decay rate for CAN

Following [62, Theorem 5], the distribution of the sojourn time T of a packet is exponentially distributed with rate ξ^* equal to the sum of the service rates at each queue minus the input rate of the class of redundant jobs, that is, $\xi^* = 2\alpha - 1$ or, equivalently, $\xi^* = (1-\varrho)/\varrho$ for $\varrho < 1$. Restating time units, the decay rate $\alpha\,\Xi^*$ of T for CAN follows from $\lambda\,\xi^* = \alpha\,\Xi^*$ with $\lambda = 2\alpha\varrho$; this yields expression (4.14) for rate Ξ^* ■

□

As required, all decay rates S^*, Σ^* and Ξ^* vanish for the maximal load admissible in the system, that is, for $\varrho = 1$, $\varrho = 1/2$ and $\varrho = 1$, respectively. Besides, these rates equal the maximal system service rate at low load,

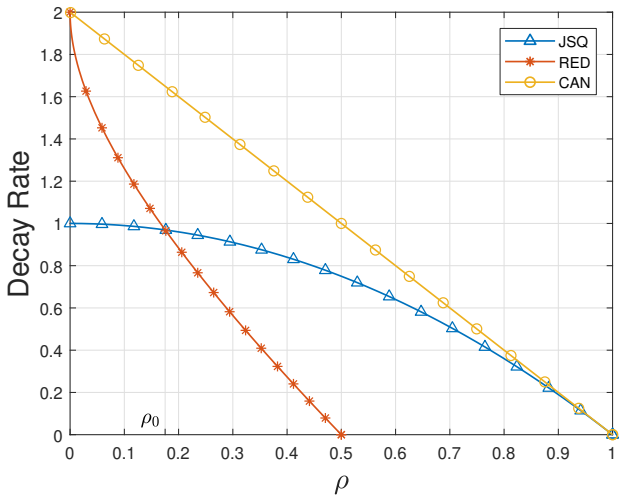


Figure 4.2: Respective decay rates S^*, Σ^* and Ξ^* for JSQ, RED and CAN allocation schemes.

namely, α, 2α and 2α, respectively. The decay rates exhibited in Proposition 1 therefore quantify the fact that both RED and CAN outperform JSQ at low load, while RED becomes poor for increasing load. Note that the decay rate for JSQ intersects with that of RED at $\varrho = \varrho_0 \approx 0.1752$. On the other hand, the performance of both JSQ and CAN become similar at high load, as illustrated in Figure 4.2. Furthermore, the scheduler can also apply the proportional policy outlined in Subsection 3.2.2 where the traffic is divided proportionally to the service rates. This scheme is equivalent to two independent M/M/1 queues, each with an arrival rate of $\lambda/2$ and service rate α. Using [64, Sect.3.2, Eq 3.31], the decay rate is $1 - \varrho$. JSQ outperforms this scheme for all values for ϱ given that it is a linear function with a maximum at $(0, \alpha)$ and a minimum at $(1, 0)$.

4.2.3 Outage Probability

Given the results of the previous Section, the outage probability $\mathbb{P}(T > t_0)$ corresponding to a given delay threshold t_0 can now be simply estimated by

$$\mathbb{P}(T > t_0) \approx \exp(-\alpha\, \Theta^* \cdot t_0), \tag{4.26}$$

with respective rate Θ^* given in Proposition 1. As outlined in Section 4.2.2.3, the distribution of T is exponential for the CAN scheme and estimation (4.26) is thus exact in this case.

Now, regarding JSQ and RED, estimation (4.26) is applicable if target delay t_0 is in the range of the distribution tail of T, that is, $t_0 \gg \mathbb{E}(T)$. In order to assess the precision of this estimation, estimate (4.26) can be compared to a numerical calculation of the distribution tail of T by expressing the latter as a contour integral in the complex plane. Recall that (M, N) denotes the vector of numbers of packets in each queue; we have the following result.

Proposition 2. ***I) Let*** G ***denote the generating function of the distribution of vector*** (M, N) ***for the JSQ scheme. Then the distribution of delay*** T ***can be expressed as***

$$\mathbb{P}(T > t) = \frac{e^{-\alpha\, t}}{2\iota\pi} \int_{|x|=r} \frac{G(x,1)}{x-1}\, e^{\alpha\, t/x}\, \mathrm{d}x \tag{4.27}$$

for all $t \geqslant 0$***, for any fixed*** $r \in \,]1, 1+2\varrho[$**.**

II) Let H ***denote the generating function of the distribution of vector*** (M, N) ***for the RED scheme. Then the distribution of delay*** T ***can be expressed as***

$$\mathbb{P}(T > t) = \frac{e^{-2\alpha\, t}}{(2\iota\pi)^2} \times \int_{|x|=r} \int_{|y|=r} \frac{H(x,y)}{(x-1)(y-1)}\, e^{\,\alpha\, t\left(\frac{1}{x}+\frac{1}{y}\right)}\, \mathrm{d}x\, \mathrm{d}y \tag{4.28}$$

for all $t \geqslant 0$ ***and any fixed*** $r \in \,]1, 1/\sqrt{2\varrho}[$**.**

Proof. We first start with the following lemma.

a) For the JSQ scheme, the distribution of sojourn time T is given by

$$\mathbb{P}(T > t) = \sum_{m \geqslant 0} P_m \cdot e^{-\alpha\, t} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!}, \qquad t \geqslant 0. \tag{4.29}$$

where $P_m = \mathbb{P}(M = m)$, $m \geqslant 0$.

b) For the RED scheme, the distribution of sojourn time T is given by

$$\mathbb{P}(T > t) = e^{-2\alpha t} \sum_{m,n \geqslant 0} \Pi_{m,n} \sum_{i=0}^{m} \frac{(\alpha t)^i}{i!} \sum_{j=0}^{n} \frac{(\alpha t)^j}{j!} \tag{4.30}$$

for all $t \geqslant 0$**.**

Proof. **a)** For all $t \geqslant 0$, the definition (4.15) of T for JSQ entails

$$\begin{aligned}\mathbb{P}(T > t) &= \mathbb{P}\left(S_1 + ... + S_M + S > t\right) \\ &= \sum_{m \geqslant 0} P_m \cdot \mathbb{P}(S_1 + ... + S_m + S > t)\end{aligned}$$

where $P_m = \mathbb{P}(M = m)$, $m \geqslant 0$. Besides, given m, the identical exponential distribution of all S_k, $1 \leqslant k \leqslant m$, and S entails that the sum $S_1 + ... + S_m + S$ has an Erlang distribution with shape parameter $m + 1$ and rate $\alpha\, t$, hence

$$\mathbb{P}(S_1 + ... + S_m + S > t) = \sum_{i=0}^{m} e^{-\alpha\, t} \frac{(\alpha\, t)^i}{i!};$$

using the latter, the above expression of $\mathbb{P}(T > t)$ reduces to (4.29).

b) Given $t \geqslant 0$, the definition (4.22) of T for RED entails

$$\mathbb{P}(T > t) = \mathbb{P}\left(S_1 + ... + S_M + S > t,\ S'_1 + ... + S'_N + S' > t\right) \tag{4.31}$$

$$= \sum_{m,n \geqslant 0} \Pi_{m,n} \times \mathbb{P}\left(S_1 + ... + S_m + S > t,\ S'_1 + ... + S'_n + S' > t\right) \tag{4.32}$$

$$= \sum_{m,n \geqslant 0} \Pi_{m,n}\, \mathbb{P}(S_1 + ... + S_m + S > t) \times \mathbb{P}(S'_1 + ... + S'_n + S' > t)$$

by the independence assumption. Besides, given m and n, the identical exponential distribution of all service times S_k, $1 \leqslant k \leqslant m$, S, and S'_ℓ, $1 \leqslant \ell \leqslant n$, S', further provides

$$\mathbb{P}(S_1 + ... + S_m + S > t) = \sum_{i=0}^{m} e^{-\alpha\, t} \frac{(\alpha\, t)^i}{i!}$$

and similarly

$$\mathbb{P}(S'_1 + ... + S'_n + S' > t) = \sum_{j=0}^{n} e^{-\alpha\, t} \frac{(\alpha\, t)^j}{j!}$$

for all $t \geqslant 0$, so that the latter expression of $\mathbb{P}(T > t)$ reads as in (4.30) ■

We now turn to the proof of Proposition 2 (in the following, the arrival rate λ is again normalized to 1).

4.2.3.1 Contour integral for JSQ

As shown in [60, Lemma, Sect.2], the function F (or G) can be analytically extended from the product $\{x \in \mathbb{C},\ |x| < 1\} \times \{y \in \mathbb{C},\ |y| < 1\}$ in $\mathbb{C}^2$ to the larger product $\{x \in \mathbb{C},\ |x| < 1\} \times \{y \in \mathbb{C},\ |y| < 1 + 1/\alpha\}$. Fix then $r \in]1, 1+1/\alpha[$; by the Cauchy formula, the generating function $x \mapsto G(x,1)$ of the marginal distribution $P_m = \mathbb{P}(M = m)$, $m \geqslant 0$, enables us to express each probability P_m as the contour integral

$$P_m = \frac{1}{2\iota\pi} \int_{|x|=r} \frac{G(x,1)}{x^{m+1}}\,\mathrm{d}x, \qquad m \in \mathbb{N},$$

on the circle $\{x \in \mathbb{C},\ |x| = r\}$ and the expression (4.29) for $\mathbb{P}(T > t)$ consequently reads

$$\mathbb{P}(T > t) = \frac{e^{-\alpha\, t}}{2\iota\pi} \int_{|x|=r} \frac{G(x,1)}{x}\,\mathrm{d}x \sum_{m\geqslant 0} \frac{1}{x^m} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!} \tag{4.33}$$

for all $t \geqslant 0$. Now setting $\xi = 1/x$ for short, we have $|\xi| < 1$ so that

$$\sum_{m\geqslant 0} \xi^m \sum_{i=0}^{m} \frac{(\alpha t)^i}{i!} = \sum_{i\geqslant 0} \frac{(\alpha t)^i}{i!} \sum_{m\geqslant i} \xi^m = \sum_{i\geqslant 0} \frac{(\alpha t)^i}{i!} \frac{\xi^i}{1-\xi}$$

hence

$$\sum_{m\geqslant 0} \xi^m \sum_{i=0}^{m} \frac{(\alpha t)^i}{i!} = \frac{e^{\alpha\, t\xi}}{1-\xi}.$$

Applying the latter to (4.33) with $\xi = 1/x$, the latter eventually reads

$$\mathbb{P}(T > t) = \frac{e^{-\alpha\, t}}{2\iota\pi} \int_{|x|=r} \frac{G(x,1)}{x-1} \exp\left(\frac{\alpha\, t}{x}\right) \mathrm{d}x \tag{4.34}$$

for all $t \geqslant 0$ and any fixed $r \in]1, 1 + 1/\alpha[$.

4.2.3.2 Contour integral for RED

Following [63, Theorem 2.2], the generating function H can be analytically extended from the product of open disks $\{x \in \mathbb{C},\ |x| < 1\} \times \{y \in \mathbb{C},\ |y| < 1\}$

in $\mathbb{C}^2$ to the larger product $\{x \in \mathbb{C},\ |x| < \sqrt{\alpha}\} \times \{y \in \mathbb{C},\ |y| < \sqrt{\alpha}\}$. Fix then $r \in]1, \sqrt{\alpha}[$; by the bi-dimensional Cauchy formula, we can then express each probability $\Pi_{m,n}$, $(m,n) \in \mathbb{N}^2$, as the double contour integral

$$\Pi_{m,n} = \frac{1}{(2\iota\pi)^2} \int_{|x|=r} \int_{|y|=r} \frac{H(x,y)}{x^{m+1} y^{n+1}}\, \mathrm{d}x\, \mathrm{d}y$$

on the product of circles $\{x \in \mathbb{C},\ |x| = r\} \times \{y \in \mathbb{C},\ |y| = r\}$ and the expression (4.30) for $\mathbb{P}(T > t)$ consequently reads

$$\mathbb{P}(T > t) = \frac{e^{-2\alpha\, t}}{(2\iota\pi)^2} \int_{|x|=r} \int_{|y|=r} \frac{H(x,y)}{xy}\, \mathrm{d}x\, \mathrm{d}y \ \times \sum_{m,n \geqslant 0} \frac{1}{x^m y^n} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!} \sum_{j=0}^{n} \frac{(\alpha\, t)^j}{j!} \tag{4.35}$$

for all $t \geqslant 0$. Now setting $\xi = 1/x$, $\eta = 1/y$ for short, we have $|\xi| < 1$, $|\eta| < 1$ so that

$$\sum_{m \geqslant 0, n \geqslant 0} \xi^m \eta^n \sum_{i=0}^{m} \frac{(\alpha t)^i}{i!} \sum_{j=0}^{n} \frac{(\alpha t)^j}{j!} \ = \sum_{i \geqslant 0, j \geqslant 0} \frac{(\alpha t)^i}{i!} \frac{(\alpha t)^j}{j!} \sum_{m \geqslant i, n \geqslant j} \xi^m \eta^n$$

with $\sum_{m \geqslant i, n \geqslant j} \xi^m \eta^n = \xi^i\, \eta^j / (1-\xi)(1-\eta)$ hence

$$\sum_{m \geqslant 0, n \geqslant 0} \xi^m \eta^n \sum_{i=0}^{m} \frac{(\alpha t)^i}{i!} \sum_{j=0}^{n} \frac{(\alpha t)^j}{j!} = \frac{e^{\alpha\, t(\xi+\eta)}}{(1-\xi)(1-\eta)}.$$

Applying the latter to (4.35) with $\xi = 1/x$ and $\eta = 1/y$, the latter eventually reduces to the integral formula (4.28) for all $t \geqslant 0$ and any fixed $r \in]1, \sqrt{\alpha}[$ (restating time units, the maximum value $\sqrt{\alpha}$ of radius r corresponds to the value $\sqrt{\alpha/\lambda} = 1/\sqrt{2\varrho}$).

□

The respective expressions for generating functions G and H invoked in Proposition 2 are detailed in Section 4.2.1.1, Equ.(4.10) and 4.2.1.2, Equ.(4.13).

The calculation of the distribution of T is thus reduced to that of simple or double contour integrals. Specifically, setting $x = r\, e^{\iota u}$, $u \in [0, 2\pi]$, in contour integral (4.27) transforms the latter into an integral on the real interval $[0, 2\pi]$. Similarly, the variable change $(x, y) \mapsto (u, v)$ with $x = r\, e^{\iota u}$,

$y = r\,e^{\iota v}$, $u, v \in [0, 2\pi]$, transforms (4.28) into a double integral on the product $[0, 2\pi] \times [0, 2\pi]$ of real intervals. For the numerical implementation of formulas (4.27) and (4.28), we chose the specific value $r = 1.01$ (recall these integrals are independent of r); as empirically observed, this choice of r guarantees numerical stability in view of the rapidly oscillating exponential terms in the integrand.

4.2.4 Average Sojourn Time

A consequence of the integral formulas is the derivation of the average sojourn time $\mathbb{E}(T) = \int_{t>0} \mathbb{P}(T > t)\,\mathrm{d}t$, namely (after exchanging the integration signs)

$$\mathbb{E}(T) = \frac{1}{\alpha}\,\frac{1}{2\iota\pi}\,\int_{|x|=r} \frac{x\,G(x,1)}{(x-1)^2}\,\mathrm{d}x \tag{4.36}$$

for JSQ, and

$$\mathbb{E}(T) = \frac{1}{\alpha}\,\frac{1}{(2\iota\pi)^2}\,\times \int_{|x|=r}\int_{|y|=r} \frac{H(x,y)}{(x-1)(y-1)}\,\frac{xy\,\mathrm{d}x\,\mathrm{d}y}{2xy-x-y} \tag{4.37}$$

for RED; while the former formula for JSQ can be alternatively represented by means of series (see Section 4.2.1, Equ.(4.12)), the mean sojourn time for RED has yet no alternative expression than the double integral (4.37). These expressions of $\mathbb{E}(T)$ can be readily exploited to assess the validity of the assumption $t_0 \gg \mathbb{E}(T)$ for the application of estimation (4.26) of the outage probability. In Figure 4.3, we plot $\mathbb{E}(T)$ in terms of ϱ; we have assumed a reserved bandwidth of 2 MHz, with a spectral efficiency of 2 bits/Hz/s, to transmit URLLC packets of size $W = 32$ bytes, so that $1/\alpha = 0.064$ ms, and a target delay for the URLLC service $t_0 = 1$ ms. We can observe that the mean waiting time for $\varrho \leqslant 0.3$ is negligible (say, $t_0 \approx 10 \times \mathbb{E}[T]$) compared to the delay threshold set to $t_0 = 1$ ms.

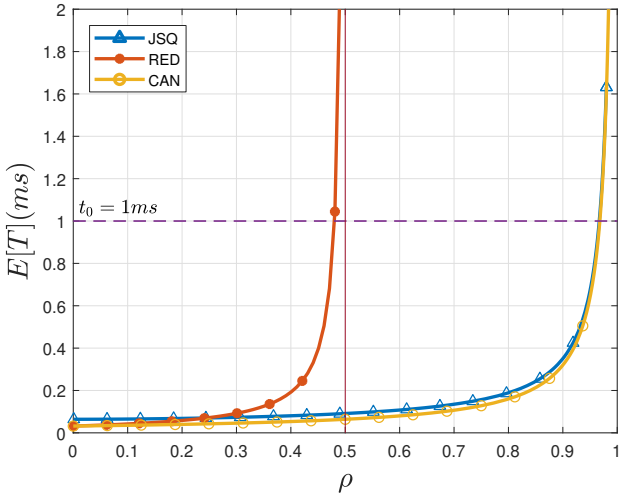


Figure 4.3: Mean sojourn time $\mathbb{E}[T]$ for JSQ, RED and CAN in terms of ϱ.

4.3 Resource dimensioning

Using the tools of previous Section 4.2, we now discuss the most suitable allocation scheme advisable for the URLLC traffic in terms of network load.

We start by assessing the validity of conclusions obtained using decay rates. Figure 4.4 plots the outage probability using both decay rates and contour integrals, with varying ϱ and fixing $1/\alpha = 0.064$ ms (corresponding to a 2 MHz system bandwidth and packets of 32 bytes). As expected, CAN outperforms JSQ and RED for all values of ϱ, while JSQ outperforms RED in the medium to high load regime only, including for a target outage of 10^{-5}. We also note that the approximation by the decay rate is, in general, good. Nevertheless, while the outage probability calculated via the decay rate for RED represents a conservative approximation to the outage calculated using the contour integral, this tendency is inverted in the JSQ case. Besides, the error for RED at very low load is caused by the numerical implementation of the double integral given in formula (4.28).

We now perform a system dimensioning exercise. Figure 4.5a shows the maximum achievable load ϱ^* as a function of the dedicated bandwidth for URLLC applications; this load ϱ^* is calculated by means of the decay rates

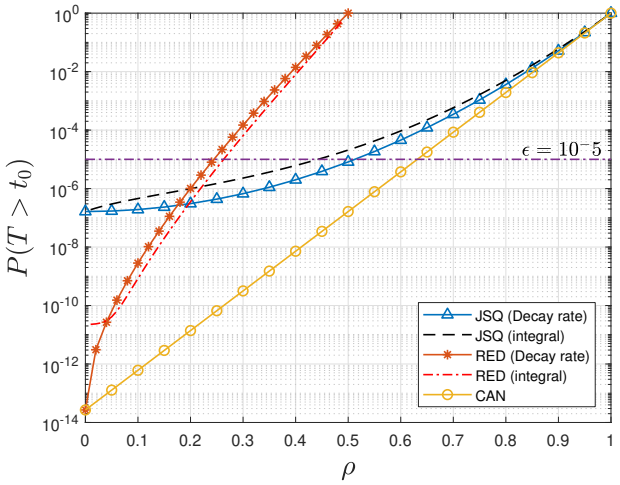
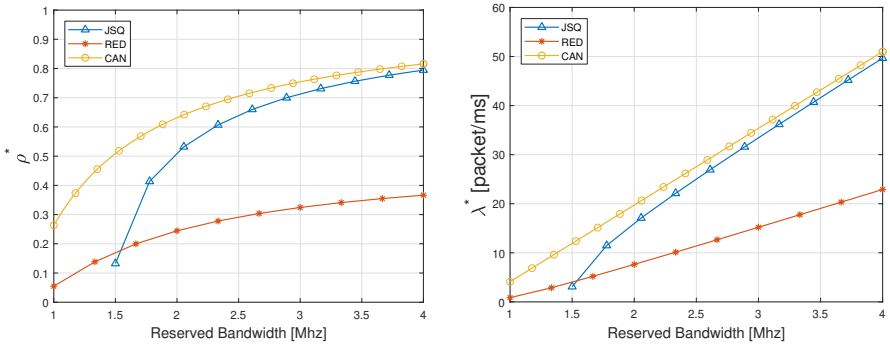


Figure 4.4: Outage probability $\mathbb{P}(T > t_0)$ with $t_0 = 1$ ms for JSQ, RED and CAN in terms of ϱ.

since the outage probability estimated by decay rates approximates well the outage probability obtained exactly through contour integral formulas. Using formula (4.26), we thus solve the equation $\mathbb{P}(T > t_0) = 10^{-5}$ for $\varrho = \varrho^*$. From the equality $\lambda = 2\alpha\varrho$, we also plot the corresponding maximum arrival rates, as displayed in Figure 4.5b. We first note that the JSQ scheme cannot reach the reliability target without reserving a capacity larger than 1.5 MHz. Furthermore, for a traffic corresponding to $\lambda \leqslant 5$ packets/ms (*i.e.* $\lambda \leqslant 1.3$ Mbit/s), the RED scheme is more adequate than JSQ as it allows one to reach the target reliability with less reserved resources. This can be justified by the fact that for these values, the maximum achievable load is less than ϱ_0, that is, $\varrho^* \leqslant \varrho_0$ (see Figure 4.2).

As an illustration, consider 50 users generating each 1 packet every 10 ms in average, resulting in a total arrival rate of 5 packets/ms; the amount of bandwidth to be reserved is then equal to 1 MHz for CAN or 1.6 MHz for RED, JSQ being unable to achieve the target performance. For a larger number of users, say 100, CAN needs 1.3 MHz, JSQ 1.7 MHz and RED 2.3 MHz of reserved bandwidth.



(a) Maximum achievable load ϱ^* (b) arrival rate λ^*

Figure 4.5: Maximum achievable load and arrival rate for JSQ, RED and CAN in terms of reserved bandwidth for a target outage 10^{-5} with $W = 32$ bytes and $t_0 = 1$ ms.

4.4 Summary

In this chapter, we have evaluated the performance of scheduling schemes for URLLC traffic in the context of 5G networks. In order to reduce the latency of packets, the redundant coverage of two frequency layers or RATs is exploited. The most straightforward scheme, named RED, always duplicates the packets on the two base stations, while the other schemes exploit the instantaneous state of the queues of the base stations and take decisions on a per-packet basis. In particular, JSQ allocates the packet to the queue with the smallest length and CAN always duplicates the packet but cancels the remaining copy upon service of the other one. We derived explicit expressions for the performance of the different schemes and show that CAN outperforms the two others in all load regimes. However, the results presented in this chapter restrict to the case where both BSs have the same bandwidth capacities. In general, there are no guarantees that this condition is valid unless both BSs belong to the same InP. Consequently, we need to extend these results to a general setting without any conditions on the base stations' bandwidth capacities.

CHAPTER 5

MULTI-CONNECTIVITY FOR URLLC SLICES: EXTENSION TO THE HETEROGENEOUS CASE

Contents

5.1 System Model 48
5.1.1 Resource allocation schemes 48
5.1.2 Queuing model 48
5.2 Performance Evaluation 49
5.2.1 Delay distribution of the JSQ scheduling scheme . 50
5.2.2 Delay distribution of the SED system 53
5.2.3 Delay distribution of the RED system 54
5.2.4 Delay distribution of the CAN system 57
5.3 Numerical experiments 58
5.3.1 Qualitative Analysis 59
5.4 Summary 61

In this chapter, we study the same system presented in chapter 4. Our objective here is to broaden the scope of the analysis to the case where the BSs have different capacities. Given that the symmetry argument we used is no longer valid, we rely on solving the equilibrium equations to derive the outage probability. In addition, we study a new policy called the Shortest Expected Delay (SED) which takes into consideration the effect of the different capacities. Consequently, we examine four packet scheduling and redundancy schemes, namely JSQ, SED, RED, and CAN, and maintain the outage probability as the performance evaluation metric.

5.1 System Model

We study a similar system to the one introduced in Figure 4.1 of a wireless system with a set of URLLC users covered by two RATs. The main difference is that the reserved bandwidths are different. The scheduler connected to both BSs picks one of the policies mentioned above. Throughout this manuscript, we refer to this configuration as the heterogeneous case.

5.1.1 Resource allocation schemes

As stated in [65], 3GPP decided that the average user plane latency for URLLC it takes to successfully deliver an application layer packet/message via the radio interface in both uplink and downlink should be 0.5 ms. Different policies can be applied when a packet belonging to a URLLC device arrives at the scheduler. We study the scheduling policies detailed in 4.1, in addition to the Shortest Expected Delay (SED) discipline. For instance, this scheduling policy assigns an arriving customer to the queue with the shortest expected delay, where delay refers to the sojourn time (waiting time plus the service time). Note that SED becomes JSQ scheduling policy in the homogeneous case under the assumption of identical service rates. In fact, JSQ is not optimal if the BSs' service rates are different since the logical choice is to join the queue with the shortest sojourn time in the system rather than the shortest queue.

5.1.2 Queuing model

We model the network architecture by two parallel queues fed by a Poisson process of URLLC packets with mean arrival rate λ, the size of packets being denoted by W (bytes). Motivated by the flexibility of the 5G NR air interface, we consider a First Come First Serve (FCFS) discipline for each queue. This means that the BS adapts its mini-slot dynamically so that one URLLC packet is served by the BS during one mini-slot [1]. Service times

[1]Note that, in cases where the amount of spectral resources is large, and the packet is small, several packets may be multiplexed in the frequency dimension in the mini-slot of smallest size (2 OFDMA symbols). Our assumption of a FCFS rule for each queue then

of packets at either queue are assumed to be mutually independent as the two BSs are supposed to use different spectrum bands and to be located in different positions, making the channels independent. As of the distribution of service times, it depends on the MCS used by the UEs, determined based on the instantaneous channel (an MCS corresponds to a service time instance, knowing that URLLC packets are generally of a constant small size). In order to ease the analysis, we assume that the resulting distribution is approximated by an exponential with rate α and β, respectively such that $\alpha \leqslant \beta$ and we set $z = \beta/\alpha$. We will show in the numerical distribution how these rates are obtained using realistic assumptions.

Given these two $M/M/1$ queues coupled by either JSQ, SED, RED or CAN discipline, we denote by M (resp. N) the number of packets in the first (resp. the second) queue. The associated equilibrium distribution of the occupancy vector (M, N) is then defined by $p_{m,n} = \mathbb{P}(M = m,\, N = n)$, $(m, n) \in \mathbb{N}^2$. Following [60], [66], [61] and, [62] respectively, this stationary distribution is then shown to exist provided that

- for JSQ, $\alpha + \beta > \lambda$, that is, $(1 + z)\alpha > \lambda$;
- for SED, $\alpha + \beta > \lambda$, $(1 + z)\alpha > \lambda$;
- for RED, $\alpha > \lambda$;
- for CAN, $\alpha + \beta > \lambda$, that is, $(1 + z)\alpha > \lambda$;

5.2 Performance Evaluation

For each of the above allocation schemes, the performance indicator is the outage probability metric $\mathbb{P}(T > t_0)$, where T is the sojourn time of a packet in the system and t_0 is the delay budget. This can be completely characterized by the distribution of T, which is difficult to obtain explicitly for JSQ, SED, and RED. In fact, it is closely related to the occupancy distribution $(p_{m,n})$. To compare the four schemes' respective performance, we first compare the delay distribution obtained using the equilibrium equations to

gives an upper bound of the performance, assuming a maximal slot size flexibility.

the simulations obtained using a discrete event simulator for JSQ and RED. Then, we compare the JSQ and SED schemes since they are very similar strategies. Throughout the remainder of the chapter, we fix $1/\alpha = 0.064$ ms (corresponding to a $B = 2$ MHz system bandwidth, spectral efficiency of $e = 2$ bits/Hz/s, and packets of size $W = 32$ bytes, so it is straightforward that $\alpha = B \times e/W$). We vary z to account for the impact of heterogeneity on the performance of the scheduling schemes.

5.2.1 Delay distribution of the JSQ scheduling scheme

We consider two M/M/1 queues ruled by the "Join the Shortest Queue" (JSQ) discipline. An arriving packet joins the shortest queue unless both queues have equal lengths, then he joins the first queue with probability $q^{'} = 1 - q$ and the second queue with probability q, where q is chosen randomly between 0 and 1 [67].

The equilibrium equations for $p_{m,n}$ formulated below are found by equating for each state the rate into and the rate out of the same state, where $\kappa = \lambda + \alpha + \beta$.

$$
\begin{aligned}
\kappa p_{m,n} &= \lambda p_{m-1,n} + \alpha p_{m+1,n} + \beta p_{m,n+1} \quad m > 0,\ n > m+1 \\
\kappa p_{n-1,n} &= \lambda p_{n-2,n} + \alpha p_{n,n} + \beta p_{n-1,n+1} + q\lambda p_{n-1,n-1} \quad m > 0,\ n = m+1 \\
\kappa p_{m,n} &= \lambda p_{m,n-1} + \alpha p_{m+1,n} + \beta p_{m,n+1} \quad n > 0,\ m > n+1 \\
\kappa p_{m,m-1} &= \lambda p_{m,m-2} + \alpha p_{m+1,m-1} + \beta p_{m,m} + q^{'}\lambda p_{m-1,m-1} \quad n > 0,\ m = n+1 \\
\kappa p_{n,n} &= \lambda(p_{n-1,n} + p_{n,n-1}) + \alpha p_{n+1,n} + \beta p_{n,n+1} \quad n > 0 \\
(\lambda+\beta)p_{0,n} &= \alpha p_{1,n} + \beta p_{0,n+1} \quad n > 1 \\
(\lambda+\beta)p_{0,1} &= q\lambda p_{0,0} + \alpha p_{1,1} + \beta p_{0,2} \\
(\lambda+\alpha)p_{m,0} &= \alpha p_{m+1,0} + \beta p_{m,1} \qquad m > 1 \\
(\lambda+\alpha)p_{1,0} &= q^{'}\lambda p_{0,0} + \alpha p_{2,0} + \beta p_{1,1} \\
\lambda p_{0,0} &= \alpha p_{1,0} + \beta p_{0,1}
\end{aligned}
\tag{5.1}
$$

We denote by S and $S^{'}$ the duration of a test job service time at BS1 and

BS2, respectively. Given the occupancy vector (M, N), the delay T of a given job is then given by

$$T = \begin{cases} S_1 + ... + S_M + S & if M < N \\ S'_1 + ... + S'_N + S' & if N < M \\ S_1 + ... + S_M + S & if M = N, \quad w.p. \quad 1-q \\ S'_1 + ... + S'_N + S' & if M = N, \quad w.p. \quad q \end{cases} \tag{5.2}$$

All random variables S_1, ..., S_M, S are mutually independent and identically distributed, with exponential distribution with mean $1/\alpha$. in the same manner, all random variables S'_1, ..., S'_N, S' are mutually independent and identically distributed, with exponential distribution with mean $1/\beta$. For all $t \geqslant 0$, the definition (5.2) of T thus entails

$$\begin{aligned} \mathbb{P}(T > t) &= \sum_{m,n \geqslant 0} p_{m,n} \cdot \mathbb{P}\left(T > t \mid M = m, N = n\right) \\ &= \sum_{n>m} p_{m,n} \cdot \mathbb{P}(S_1 + ... + S_m + S > t) + \sum_{m<n} p_{m,n} \cdot \mathbb{P}(S'_1 + ... + S'_m + S' > t) + \\ &\quad q' \sum_{m \geqslant 0} p_{m,m} \cdot \mathbb{P}(S_1 + ... + S_m + S > t) + q \sum_{n \geqslant 0} p_{n,n} \cdot \mathbb{P}(S'_1 + ... + S'_n + S' > t) \end{aligned}$$

Besides, given m and n, the identical exponential distribution of all variables S_i, $1 \leqslant i \leqslant m$, and S provides

$$\mathbb{P}(S_1 + ... + S_m + S > t) = \sum_{i=0}^{m} e^{-\alpha\, t} \frac{(\alpha\, t)^i}{i!} \tag{5.3}$$

and similarly for all variables and S'_j, $1 \leqslant j \leqslant n$ and S',

$$\mathbb{P}(S'_1 + ... + S'_n + S' > t) = \sum_{j=0}^{n} e^{-\beta\, t} \frac{(\beta\, t)^j}{j!} \tag{5.4}$$

So that the latter expression of $\mathbb{P}(T > t)$ further reads

$$\mathbb{P}(T > t) = \sum_{m=0}^{\infty} \sum_{n=m+1}^{\infty} p_{m,n} \cdot e^{-\alpha\, t} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!} + \sum_{n=0}^{\infty} \sum_{m=n+1}^{\infty} p_{m,n} \cdot e^{-\beta\, t} \sum_{j=0}^{n} \frac{(\beta\, t)^j}{j!} + q^{'} \cdot \sum_{m \geqslant 0} p_{m,m} \cdot e^{-\alpha\, t} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!} + q \cdot \sum_{n \geqslant 0} p_{n,n} \cdot e^{-\beta\, t} \sum_{j=0}^{n} \frac{(\beta\, t)^j}{j!}, \qquad t \geqslant 0. \tag{5.5}$$

Equation 5.5 requires the resolution of the equilibrium equations. We solve the system of equilibrium equations by adding blocking equations. We consider that all packets arriving while there are L waiting packets as lost. Thus, we can write the blocking equations as follows:

$$\begin{aligned}
\kappa p_{m,L} &= \lambda p_{m-1,L} + \alpha p_{m+1,L} \qquad 0 < m < L-1 \\
\kappa p_{L-1,L} &= \lambda p_{L-2,L} + \alpha p_{L,L} + q\lambda p_{L-1,L-1} \\
\kappa p_{L,n} &= \lambda p_{L,n-1} + \beta p_{L,n+1} \qquad 0 < m < L-1 \\
\kappa p_{L,L-1} &= \lambda p_{L,L-2} + \beta p_{L,L} + (1-q)\lambda p_{L-1,L-1} \\
(\alpha+\beta) p_{L,L} &= \lambda (p_{L-1,L} + p_{L,L-1}) \\
(\lambda+\beta) p_{0,L} &= \alpha p_{1,L} \\
(\lambda+\alpha) p_{L,0} &= \beta p_{L,1}
\end{aligned} \tag{5.6}$$

Knowing that the sojourn time of a tagged job at BS1 or BS2 finding k jobs in the system follows an Erlang distribution with shape $k+1$ and rate α or β respectively and keeping in mind that the sojourn time should be less than 0.5 to avoid an outage event. Consequently, we have $(k+1)/\beta \leqslant (k+1)/\alpha < 0.5$ ms. Thus $k < 6.8125$ for BS1 and $k < 14.625$ for BS2. In the numerical application, we choose $L = 20$. Since our goal is to quantify the outage and compare it to a system without blocking, we pick $L > k$.

We solve the linear system described by equations 5.1 and 5.6 to find $\{p_{m,n}\}$, $(m,n) \in [0..L]^2$ and use equation 5.5 to plot an approximation of $\mathbb{P}(T > t_0)$. In Figure 5.1, we compare the outage probability obtained using equilibrium equations with its counterpart found using discrete-event simulations for different values of z. We focus on traffic regimes where the outage probability is lower than 10^{-5}, defined by 3GPP as a key performance indicator for a plethora of URLLC-centered services [68, 65].

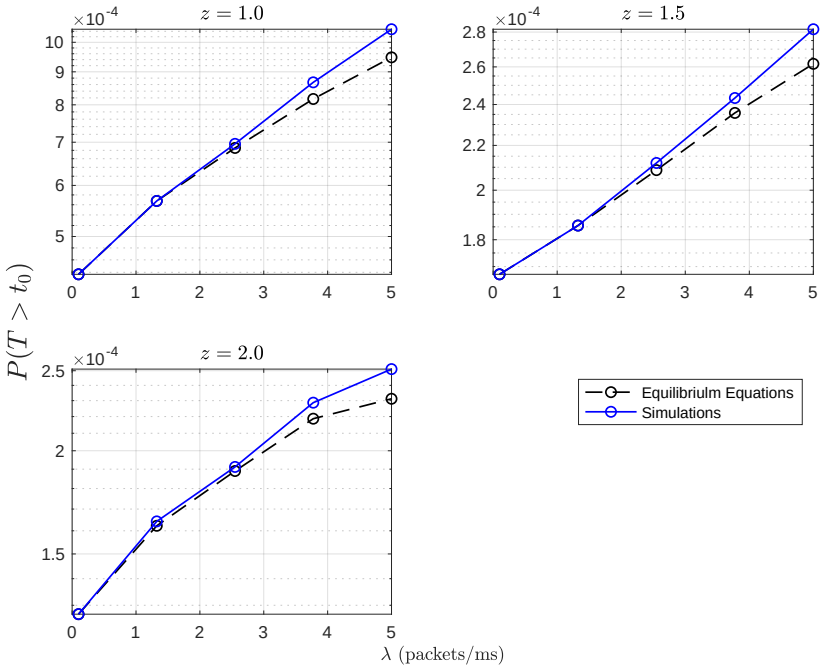


Figure 5.1: Outage probability for the JSQ scheme using equilibrium equations and simulations with $t_0 = 0.5$ ms

5.2.2 Delay distribution of the SED system

This section analyzes the performance of a system with two servers under the shortest expected delay (SED) scheduling scheme. This policy steers an arriving packet to the queue with the shortest expected delay, where delay refers to the waiting time plus the service time.

Let m and n be the number of customers in the first and second queue, respectively, including a possible customer in service. For an arriving packet, the expected delay in the first queue is $(m+1)/\alpha$ and in the second queue is $(n+1)/\beta$. The SED scheduling scheme assigns an arriving packet to queue 1 if $(m+1)/\alpha < (n+1)/\beta$ and to queue 2 if $(m+1)/\alpha > (n+1)/\beta$. When the expected delays in both queues are equal, i.e., $\beta(m+1) = \alpha(n+1)$, the arriving customer joins queue 1 with probability q and queue 2 with

probability $1 - q$.

In Figure 5.2, we plot the outage probability using discrete-event simulations. Similar to the JSQ case, we notice that the outage probability decreases with increasing values of z and increases with rising arrival rates.

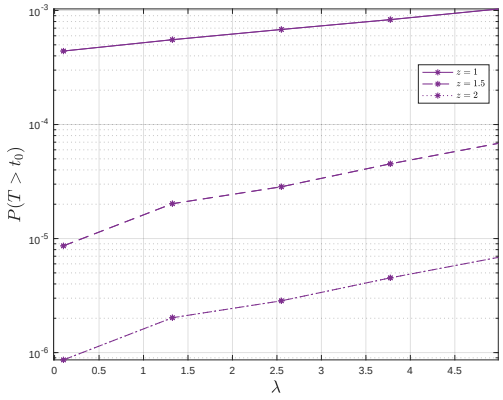


Figure 5.2: Outage probability for the SED scheme using simulations with $t_0 = 0.5$ ms

5.2.3 Delay distribution of the RED system

The systems mentioned above of two coupled queues can be compared to that of two parallel FCFS queues created by arrivals with two demands, as analyzed in [61] and [63]. The incoming packet is duplicated and sent to both queues, where each copy is served independently.

In a similar manner to the JSQ scheme, we formulate the equilibrium equations for $p_{m,n}$ by equating for each state the rate into and the rate out of the same state, where $\kappa = \lambda + \alpha + \beta$.

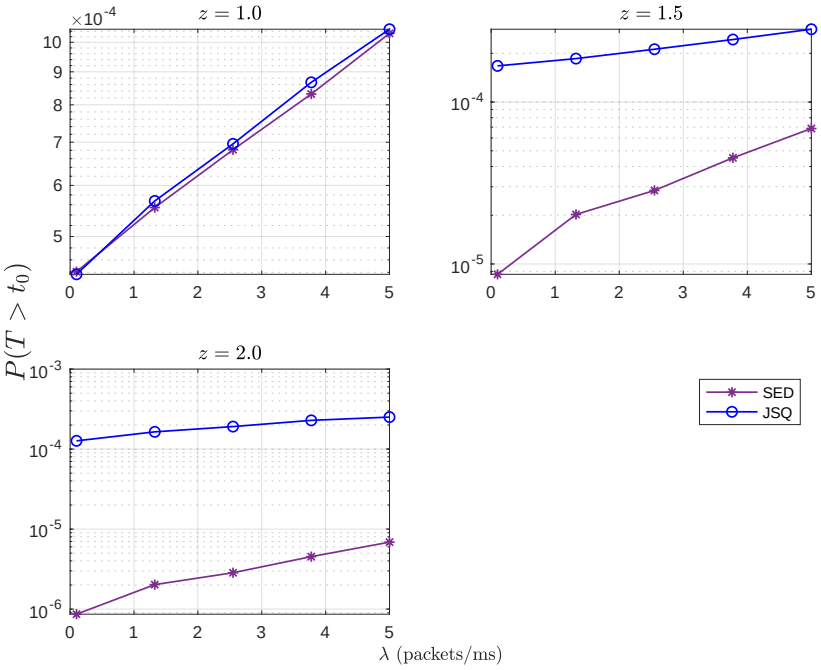


Figure 5.3: Outage probability comparison of the JSQ and SED schemes using simulations with $t_0 = 0.5$ ms

$$\begin{aligned} \kappa p_{m,n} &= \lambda p_{m-1,n-1} + \alpha p_{m+1,n} + \beta p_{m,n+1} \quad m > 0, n > 0 \\ (\lambda + \beta) p_{0,n} &= \alpha p_{1,n} + \beta p_{0,n+1} \quad n > 0 \\ (\lambda + \alpha) p_{m,0} &= \alpha p_{m+1,0} + \beta p_{m,1} \quad m > 0 \\ \lambda p_{0,0} &= \alpha p_{1,0} + \beta p_{0,1} \end{aligned} \tag{5.7}$$

Given the occupancy vector (M, N), the delay T of a given job is given by the minimum

$$T = \min \left(S_1 + ... + S_M + S,\ S'_1 + ... + S'_N + S' \right) \tag{5.8}$$

where all random variables S_1, ..., S_M, S and S'_1, ..., S'_N, S' are mutually independent and identically distributed, with exponential distribution with mean $1/\alpha$ and $1/\beta$, respectively. Given $t \geqslant 0$, the definition (5.8) of T entails

$$\begin{aligned}\mathbb{P}(T>t) &= \mathbb{P}\left(S_1+...+S_M+S>t,\ S'_1+...+S'_N+S'>t\right)\\ &= \sum_{m,n\geqslant 0} p_{m,n}\,\mathbb{P}\left(S_1+...+S_m+S>t,\ S'_1+...+S'_n+S'>t\right)\\ &= \sum_{m,n\geqslant 0} p_{m,n}\,\mathbb{P}(S_1+...+S_m+S>t)\times\mathbb{P}(S'_1+...+S'_n+S'>t)\end{aligned}$$

by the independence assumption. Besides, using equations 5.3 and 5.4, the latter expression of $\mathbb{P}(T>t)$ further reads

$$\mathbb{P}(T>t) = \sum_{m,n\geqslant 0} p_{m,n}\cdot e^{-(\alpha+\beta)\,t}\sum_{i=0}^{m}\frac{(\alpha\,t)^i}{i!}\sum_{j=0}^{n}\frac{(\beta\,t)^j}{j!},\qquad t\geqslant 0. \tag{5.9}$$

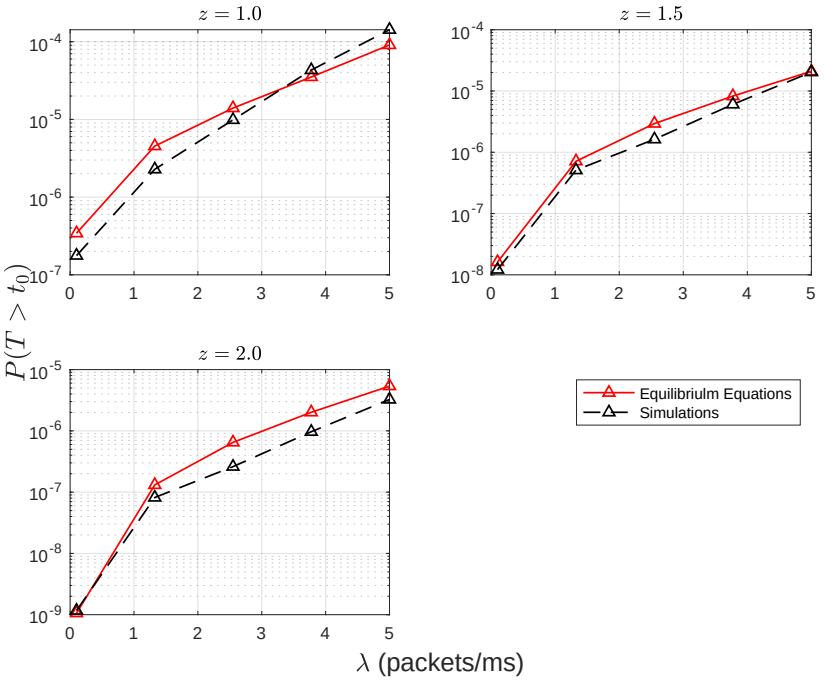


Figure 5.4: Outage probability for the RED scheme using equilibrium equations and simulations with $t_0 = 0.5$ ms

Applying the same reasoning as the JSQ case, we write the blocking equations as follows:

$$
\begin{aligned}
\kappa p_{L,n} &= \lambda p_{L-1,n-1} + \beta p_{L,n+1} \quad 0 < n < L \\
\kappa p_{m,L} &= \lambda p_{m-1,L-1} + \alpha p_{m+1,L} \quad 0 < m < L \\
(\lambda + \beta) p_{0,L} &= \alpha p_{1,L} \\
(\lambda + \alpha) p_{L,0} &= \beta p_{L,1} \\
(\alpha + \beta) p_{L,L} &= p_{L-1,L-1}
\end{aligned}
\tag{5.10}
$$

We solve the linear system described by equations 5.7 and 5.10 to find $\{p_{m,n}\}$, $(m,n) \in [0..L]^2$ with $L = 15$. We use equation 5.9 to plot an approximation of $\mathbb{P}(T > t_0)$. Figure 5.4 compares the outage probability obtained using equilibrium equations and using discrete-event simulations for different values of z. The outage probability found using the equilibrium equations for a system with blocking represents a lower bound to the outage probability for our system. For the next section, we keep the outage probability found using the equilibrium equations due to the fast computational time compared to simulations.

5.2.4 Delay distribution of the CAN system

Following [62, Theorem 5], the distribution of the sojourn time T of a packet is exponentially distributed with rate ξ^* equal to the sum of the service rates at each queue minus the input rate of the class of redundant jobs, that is, $\xi^* = \alpha + \beta - \lambda$ or, equivalently, $\xi^* = (1+z)\alpha - \lambda$ for $\lambda < \alpha + \beta$. Thus, the outage probability $\mathbb{P}(T > t_0)$ corresponding the delay threshold t_0 can be simply expressed as

$$
\mathbb{P}(T > t_0) = \exp(-\xi^* t_0) \tag{5.11}
$$

Figure 5.5 plots the outage probability as a function of λ and z using equation 5.11. Likewise, we see that the outage probability decreases with increasing values of z and increases with growing arrival rates.

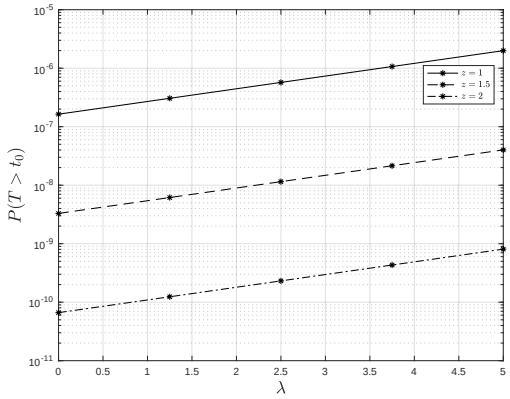


Figure 5.5: Outage probability $\mathbb{P}(T > t_0)$ with $t_0 = 0.5$ ms for CAN in terms of λ for different values of z.

5.3 Numerical experiments

In this section, we assess the performance of the schemes mentioned above by comparing the outage probability for different values of λ and z. As mentioned before, we fix $1/\alpha = 0.064$ ms which corresponds to a 2 MHz system bandwidth, spectral efficiency of 2 bits/Hz/s and packet size of 32 bytes and vary z. Figure 5.6 plots the outage probability using the equilibrium equations for RED, simulations for JSQ and SED, and equation 5.11 for CAN for a target latency of $t_0 = 0.5$ ms.

As expected, CAN outperforms JSQ, SED, and RED for all values of λ and z. While the outage probability obtained using JSQ coincides with the one achieved SED for $z = 1$, the latter outperforms JSQ for $z = 1.5$ and $z = 2$ for all values of λ. Additionally, RED outperform SED for low arrival rates (for λ less or equal than $\lambda^* = 8$, $\lambda^* = 10$ and $\lambda^* = 11$ packets/ms for $z = 1$, $z = 1.5$ and $z = 2$, respectively). λ^* represents the intersection point between SED and RED. Note that for a target outage 10^{-5}, RED and CAN are more suitable.

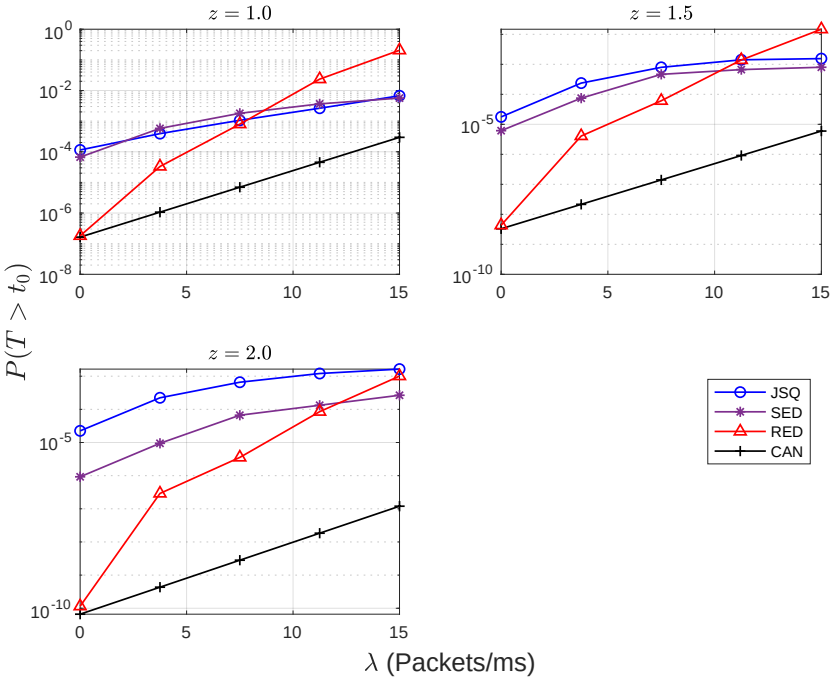


Figure 5.6: Outage probability $\mathbb{P}(T > t_0)$ with $t_0 = 0.5$ ms for JSQ, SED, RED and CAN in terms of λ for different values of z.

5.3.1 Qualitative Analysis

The decision to deploy one or another scheduling scheme depends on the achieved performance and the feasibility of the different solutions. We omit JSQ from our analysis given that it is equivalent to SED for the homogeneous case and substandard for the heterogeneous case. Although the CAN allocation scheme displays good results compared to the RED and SED policies, it remains intricate to implement. First, the BS should notify the RAN NSSMF upon each packet completion. Then, the RAN NSSMF should forward this information to the other BS to remove the remaining copy from the queue. This paradigm requires seamless knowledge about the system and can be achieved if the BSs are co-located. If not, the delay prompted by the communication links may destroy the advantage of CAN and degrade it to a RED scheme.

When BSs are connected through a limited backhaul, the NSSMF selection reduces to RED and SED schemes. Note that RED does not necessitate that BSs share any coordination or control data. Conversely, SED needs to estimate both BSs' instantaneous load upon packet arrival to steer it appropriately, introducing a limited communication overhead on the backhaul. Therefore, an efficient policy consists in dynamically altering the allocation scheme depending on the arrival rate.

As we can see in Figure 5.6, RED is favorable for low arrival rates up to λ^*. If the instantaneous system load exceeds λ^*, a shift in resource allocation schemes is thus needed to respect the URLLC reliability requirements.

We now examine the feasibility of the different schemes in the uplink. Recall that we argue that the NSSMF has access to information about the instantaneous load at each BS. It can decide the strategy to steer the downlink packets via the backhaul/fronthaul to the suitable BS. As for the uplink, the process differs significantly. The end-users generate packets that send scheduling requests to the different BSs. The latter replies with a grant indicating the time/frequency resource to be used for the packet. While RED is directly applicable in this case, SED and CAN need some additional signaling that can be specified as follows:

- For SED, when the BS receives the scheduling request, it forwards it to the NSSMF that indicates whether it has to issue a scheduling grant for the user;
- For CAN, both BSs issue a scheduling grant, as if a RED scheme were applied. However, when a BS finishes serving a packet, it signals it to the other BS. The latter may then delete the scheduling grant and reschedule another packet on the liberated resource, provided it has the necessary time and flexibility. Such a fast rescheduling is possible in 5G NR due to the dynamic in-resource scheduling feature, where an uplink scheduling grant may accompany the data intended for a user in the downlink [69].

5.4 Summary

In this chapter, we extended the performance model of chapter 4 to the heterogeneous case. We considered a new scheduling scheme called SED that sends the packet to the queue with the shortest expected delay. We derived explicit expressions for the performance of the different schemes and show that CAN outperforms the other policies for all arrival rates. However, CAN needs strict coordination between the two BSs. In the absence of such coordination, RED is preferred at low arrival rates while SED is better otherwise. In the next chapter, we will study the impact of architectural options discussed in chapter 3 on the different schemes.

CHAPTER 6

ASSESSING ARCHITECTURE IMPACT ON SCHEDULING POLICIES

Contents

6.1 Simulations for systems serving URLLC slices . 63
6.1.1 System model 63
6.2 Performance evaluation 65
6.2.1 Bandwidth Reservation Case for URLLC Slice . . 65
6.2.2 Coexistence of eMBB and URLLC Slices 67
6.3 Case Study: a smart factory served by three BSs 70
6.4 Summary 71

In this chapter, we aim at evaluating the scheduling schemes presented in the thesis in the presence of imperfections introduced by the architectural scenarios outlined in section 3.2. Our main goal is to examine the performance of RED and two variants of JSQ or SED scheduling policies depending on BSs capacities. The analysis is based on simulations in order to be able to take into consideration additional system parameters such as information delay. We also consider a more realistic service discipline where the service time depends on the MCS and subsequently on the radio conditions, whereas in the analytical modeling, it is drawn from an exponential distribution. We also extend the analysis to the case of three base stations and assess the impact of URLLC scheduling on eMBB performance. We discard the CAN scheme from our analysis due to the fact that it requires control plane information to be shared between base stations for every served packet, which

is counter-intuitive if we consider that once we issue a scheduling grant, the scheduler cannot free the reserved resource for another user, thus rendering the CAN scheme advantages obsolete.

6.1 Simulations for systems serving URLLC slices

In this section, we simulate the system presented in Figure 3.2 in three different scenarios. First, we examine the system with only URLLC packets through a resource reservation scenario for URLLC slices. Then, we consider the case where eMBB and URLLC slices share the same resources in two separate settings. We aim to gather quantitative and qualitative insights on architectural consideration's impact on delivering the stringent latency requirement of URLLC services.

6.1.1 System model

We consider a wireless system with a set of URLLC and eMBB users located within a smart factory served by two RATs with bandwidth B_1 and B_2, respectively. URLLC packets are steered with regard to the network architecture and the placement of resource management entities (see Figures 3.3 and 3.4). Driven by the 5G NR air interface's flexibility, URLLC packets are served on a mini-slot basis of 2 OFDM symbols, whereas eMBB packets are served on a legacy 1 ms TTI [70]. Service times of URLLC and eMBB packets depend on the used modulation and coding scheme. The latter differs from one user to another, depending on its average radio conditions.

We model the network architecture by two parallel queues fed by a Poisson process of URLLC packets of size W with mean arrival rate per user denoted μ. Due to heterogeneous radio conditions, the Modulation and Coding Schemes (MCSs) of users are different. Let $\mathcal{S}$ be the set of spectral efficiencies associated to the different MCS, and let p_s be the probability of having spectral efficiency $s \in \mathcal{S}$. The service time of the i-th URLLC packet at BS j is $1/\alpha_{j,i}$ where

$$\alpha_{j,i} = \frac{B_j \times X_{j,i}}{W},$$

and $X_{j,i} \in \mathcal{S}$ is the efficiency of the MCS used by packet i on BS j.

When URLLC and eMBB slices share the same resources, we assume each BS serves a set of eMBB users separately. Two independent Poisson processes generate the eMBB packets of size $Z >> W$, with arrival rates λ_1 and λ_2. Hence, the service time of eMBB packet k at BS j is $1/\psi_{j,k}$ where:

$$\psi_{j,k} = \frac{B_j \times Y_{j,k}}{Z},$$

where $Y_{j,k}$ is the spectral efficiency for eMBB packet k on BS j. We denote by ϱ_j the eMBB traffic load at BS j, defined as:

$$\rho_j = \frac{\lambda_j}{\hat{\psi}_j},$$

where $\hat{\psi}_j$ is the average service rate for eMBB packets at BS j.

We study two different policies based on the architectural options discussed above:

1. **The decision in a shared RAN NSSMF:** When the scheduling decision is taken at the RAN NSSMF level, the packet steering policy depends on the base stations' load. This scheme consists of sending the incoming URLLC packet to the queue with the smallest number of waiting packets. We consider two practical variants. The first assumes that the NSSMF knows the instantaneous load with a minimal control plane delay, set to 100 μs. The second case takes into account the control plane signaling delay equals $1ms$ in the numerical application. In other terms, the NSSMF relies on information reports sent by the BSs some time ago to make its decision. In both cases, we apply the JSQ and SED schemes in the homogeneous and heterogeneous configurations, respectively.

2. **The decision in a far NSMF:** When the instantaneous load is not available as the decision is taken at the NSMF level, we consider RED

as a possible resource allocation scheme. This scheme does not require any prior knowledge of the radio access channel. Therefore, it does not entail substantial control plane information.

6.2 Performance evaluation

In the following, we evaluate the outage probability of URLLC traffic originating from the above allocation schemes using Monte Carlo simulations. The outage probability is defined as the probability that the packets' latency exceeds a predefined delay budget set to 0.5 ms.

Simulation parameters	**Value**
URLLC packet size	32 bytes
eMBB packet size	1500 bytes
Control plane reports	100 μs, 1 ms
Latency threshold	0.5 ms
URLLC packet generation per user	100 packets/s
URLLC Spectral efficiency	$\{1, 1.5, 2, 2.5\}$ bits/Hz/s
eMBB spectral efficiency	9 bits/Hz/s [71]

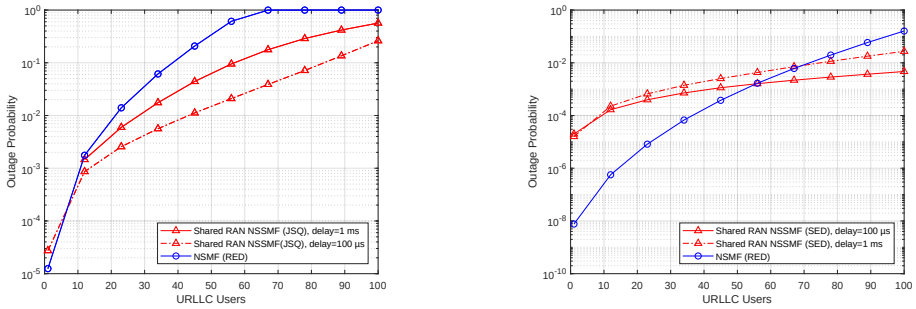
Table 6.1: Parameters for performance evaluation

We study three distinct scenarios, each in two separate settings: the homogeneous case (*i.e.,* $B_1 = B_2$) and the heterogeneous case with dissimilar bandwidths at the BSs. Table 6.1 summarizes the system setting for performance evaluation.

6.2.1 Bandwidth Reservation Case for URLLC Slice

First, we study the impact of slicing architecture on URLLC traffic. In this scenario, we reserve a sub-band for URLLC traffic on each BS to achieve hard isolation with the eMBB traffic. We examine the homogeneous and the heterogeneous setting where we assume a reserved bandwidth of $(B_1, B_2) = (1, 1)$ MHz and $(B_1, B_2) = (2, 1)$ MHz, respectively. The URLLC packets' mean arrival rate per user is set to $\mu = 100$ packets/s. Figure 6.1 shows

the URLLC traffic's outage probability stemming from the different policies while increasing the number of URLLC users in the smart factory.



(a) (b)

Figure 6.1: Outage probability in the case of bandwidth reservation for URLLC packets. **(a)** $B_1 = B_2 = 1$ MHz; **(b)** $B_1 = 2, B_2 = 1$ MHz

We observe two regimes through a comprehensive look at the proposed allocation schemes' performance, each giving an advantage for one of the architectural options. Figure 6.1a shows that redundancy displays the best performance for a restricted number of URLLC users in the homogeneous case. In the medium to high load regimes, placing the intelligence at a shared RAN NSSMF with reduced control delay has an advantage over the far NSMF entity's management. Besides, a large control delay worsens the performance of the shared RAN NSSMF policy. Note that the latter outperforms the redundancy scheme in high load regimes since it circumvents overloading. However, high load regimes are not suitable for meeting URLLC QoS requirements, where a very low outage probability is sought, in the order of 10^{-6} to 10^{-5}.

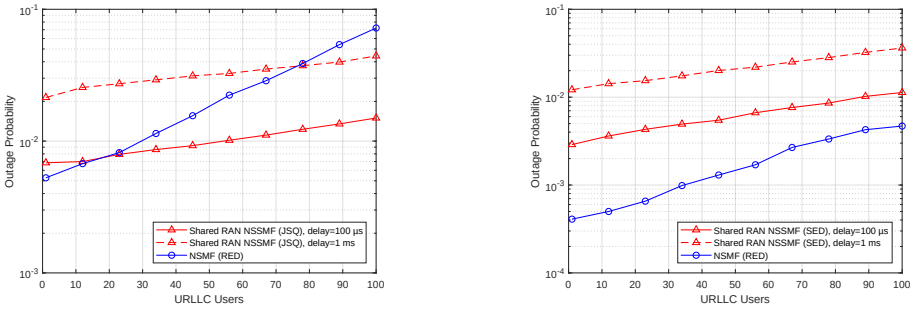
In the heterogeneous case (see Figure 6.1b), we notice that the redundancy policy profits from the asymmetric reserved bandwidth at the BSs compared to the previous case. This can be explained by the fact that duplicated URLLC packets undergo almost the same service time. Consequently, the minimum sojourn time at the system is not significantly reduced. Thus, only packet duplication can achieve the target QoS in this low load regime without the need for exhaustive cooperation.

6.2.2 Coexistence of eMBB and URLLC Slices

We now move to a setting where URLLC and eMBB slices share the same resources. We exploit the overall bandwidth without reserving a fixed band for URLLC traffic. Again, we study the homogeneous and the heterogeneous case where the overall bandwidth is $(B_1, B_2) = (10, 10)$ MHz and $(B_1, B_2) = (20, 10)$ MHz, respectively. Our objective is twofold. First, we aim to study the impact of eMBB and URLLC multiplexing on the URLLC performance, and second, we aim at reinspecting the role of URLLC slice management function placement.

6.2.2.1 Variable URLLC traffic with fixed eMBB traffic

To obtain insights on the impact of coexistence between eMBB and URLLC traffic, we gradually increase URLLC users' number while maintaining the eMBB load at each BS at $\rho_1 = \rho_2 = 0.7$. Like the previous case, we set the URLLC packets' mean arrival rate per user to $\mu = 100$ packets/s. In Figure 6.2, we plot the outage probability for URLLC packets. We first remark that the outage probability has higher values than the previous case (separated URLLC/eMBB) since the URLLC packets compete for radio resources with large eMBB packets. Ultra-reliability is thus very difficult to achieve when there is no strict resource reservation for URLLC traffic.

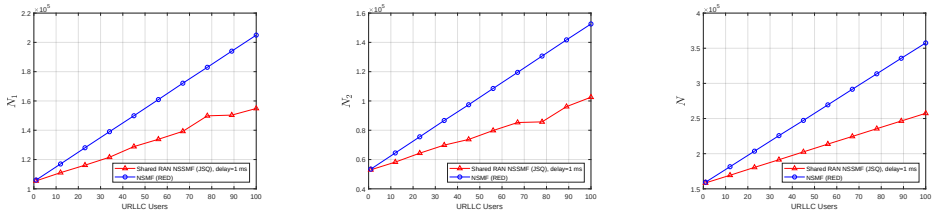


(a) (b)

Figure 6.2: Outage probability in the case of fixed eMBB and variable URLLC traffic without bandwidth reservation.**(a)** $B_1 = B_2 = 10$ MHz; **(b)** $B_1 = 20, B_2 = 10$ MHz

We now have a deeper look at the performance of the different URLLC slice management policies. Figure 6.2a shows that the NSMF redundancy degrades the performance for mid to high loads but is essential for achieving high reliability. Indeed, even if it increases the load, redundancy increases the chance that duplicated packets have access to the queue with minimal awaiting eMBB packets, thus reducing the URLLC packets' sojourn time. However, tight coordination at the RAN NSSMF level offers good results but is still outperformed in low load regimes by packet duplication. Therefore, it is recommended to design a dynamic strategy where we move from an NSMF redundancy to a shared NSSMF policy based on the number of URLLC users in the factory.

Figure 6.2b shows the outage probability in the heterogeneous configuration. The performance trend is similar to that of the homogeneous counterpart. The difference is that we need not change our policy dynamically since the NSMF redundancy policy outperforms the other policies for all traffic load regimes.



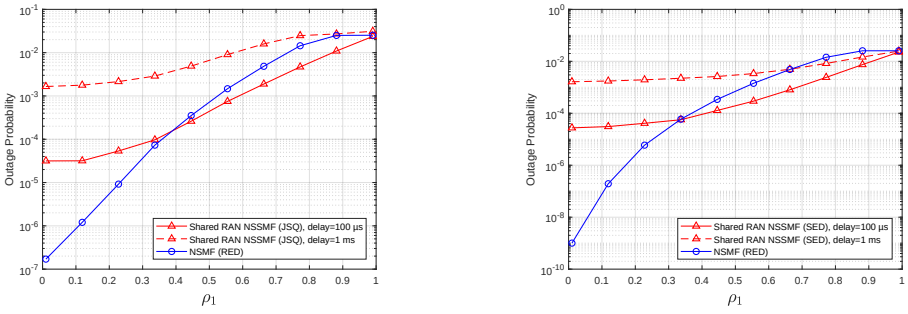
(a) (b) (c)

Figure 6.3: Number of packets served in BS1, BS2 and the system, respectively with $B_1 = B_2 = 10$ MHz.**(a)** BS1 ; **(b)** BS2 ; **(c)** System

We denote by N_1, N_2 and N the number of packets served in BS1, BS2, and the overall system (*i.e.* the sum of packets served in BS 1 and 2). It is essential to point out that applying the redundancy-based scheme in both the uplink and the downlink instigates an over-utilization of resources that we quantify in Figure 6.3. Hence, we can clearly see that respecting the latency requirement of URLLC use cases degrades the eMBB users' rates. This degradation can also be caused by scheduling URLLC packets over mini-slot while puncturing eMBB transmission [72].

6.2.2.2 Variable eMBB traffic with fixed URLLC traffic

We investigate another setting where we have a fixed number of URLLC users set to 100. The URLLC packets' mean arrival rate per user is set to $\mu = 100$ packets/s. We vary the eMBB load at one BS while maintaining the traffic load with $\rho_2 = 0.5$ at the other.



(a) (b)

Figure 6.4: Outage probability in the case of variable eMBB at BS1, fixed eMBB traffic at BS2, and fix number of URLLC users.**(a)** $B_1 = B_2 = 10$ MHz; **(b)** $B_1 = 20, B_2 = 10$ MHz

We first note that the outage probability for the heterogeneous configuration shows a similar performance trend to the homogeneous case. For instance, the NSMF redundancy policy presents a lower outage probability compared to the other scheduling policies, up to BS1 load $\rho_1 \approx 0.4$ and $\rho_1 \approx 0.33$, for the homogeneous and the heterogeneous case, respectively (see Figures 6.4a and 6.4b). These values represent a threshold for designing a dynamic strategy, where we change the scheduling from the NSMF redundancy to the shared NSSMF. Again, The control plane signalization degrades the shared NSMF performance.

Figure 6.5 shows the eMBB throughput as a function of the eMBB traffic load at BS1. Our goal is to quantify the impact of the scheduling policies on the performance of eMBB services. We can see that, although redundancy is vital to achieving URLLC requirements in terms of low outage probabilities, as shown above, it leads to the degradation of the eMBB throughput due to the inefficient use of resources. To summarize, when the slice scheduling functions are placed far from the BSs, introducing a delayed decision,

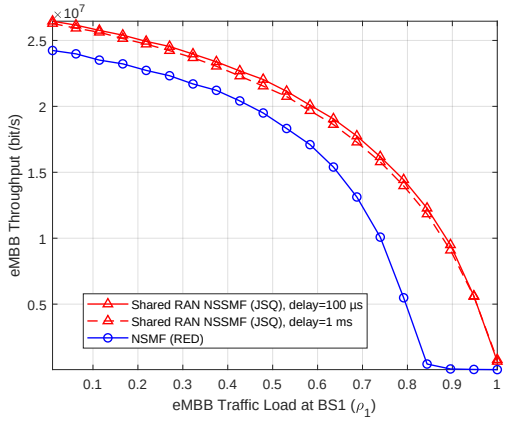


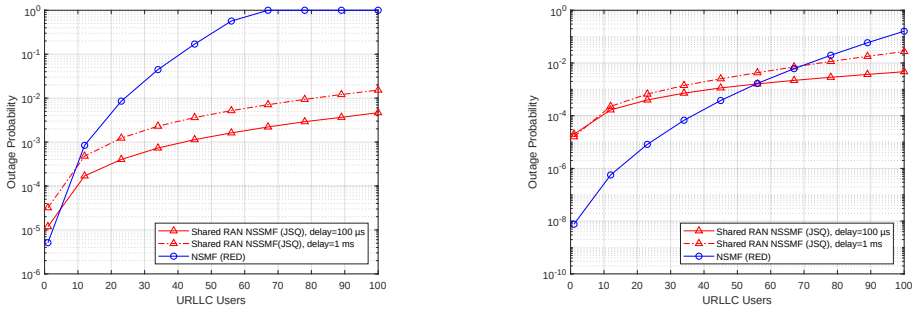
Figure 6.5: eMBB throughput in the case of variable eMBB traffic at BS1, fixed eMBB traffic at BS2, and fix number of URLLC users, where $B_1 = B_2 = 10$ MHz.

the stringent delay requirements of URLLC cannot be achieved with a per-packet scheduling policy, and a systematic redundancy is needed, leading to inefficiencies in resource usage.

6.3 Case Study: a smart factory served by three BSs

We corroborate the results shown in section 6.1 by simulating the case where 3 BSs are co-located and serve a set of URLLC users. We denote by B_3 the reserved bandwidth at BS3. Likewise, we reserve a sub-band for URLLC traffic on each BS set $B_1 = B_2 = B_3 = 1$ MHz for the homogeneous case, and $B_1 = 2, B_2 = B_3 = 1$ MHz for the heterogeneous case.

Figure 6.6 displays the URLLC traffic's outage probability originating from the different policies while raising the number of URLLC users in the smart factory. Similar observations to the two BS case can be made. In particular, the shared RAN NSSMF case, as it manages load instantaneously, outperforms the NSMF case. However, this advantage vanishes when the delay increases. On the other hand, the systematic redundancy case outperforms



(a) (b)

Figure 6.6: Outage probability in the case of bandwidth reservation for URLLC packets. **(a)** $B_1 = B_2 = B_3 = 1$ MHz; **(b)** $B_1 = 2, B_2 = B_3 = 1$ MHz

the remaining schemes for low load, but its performance degrades for high load regime. The same tendency is observed for the heterogeneous case. The difference is that the amount of resources is more significant, leading to a switch towards the high load regime.

6.4 Summary

This chapter offered a simulation-based analysis of some of the scheduling policies studied analytically in chapters 4 and 5. We aimed to explore the network slicing architecture's impact on the placement of the management entities involved in resource allocation and traffic steering decisions in both the homogenous and the heterogeneous case, using the same framework. In particular, we studied two architectural options, the first with loose coupling, where the scheduling policy was determined within the NSMF by the vertical where RED is applied. The second case used tight coupling where the scheduling decision was taken at the NSSMF level by the mobile network operator who owns the different BSs. On the one hand, we utilize JSQ while varying the control information transmission delay for the homogenous case. On the other hand, we assigned SED to the heterogeneous case using the same values for control delays. Both methods can only be used when information about the number of waiting packets or the average sojourn time is

accessible.

The simulation results are consistent with the analytical results. For instance, systematic redundancy, which does not require any tight coupling between BSs, was crucial to achieving a low outage probability for a low URLLC load regime. The results also show that while tight coupling was beneficial for the system in higher traffic loads, it lost its efficiency rapidly when information about each cell's load arrived with a delay because of the outdated scheduling decision. This effect occurs when the slice scheduling functions are placed far from the BSs, introducing a control plane delay. Similarly, RED remains crucial to delivering low outage probabilities in the coexistence scenario, leading to ineffective resource exploitation impacting eMBB traffic performance. We showed that the same performance pattern could be observed in case three BSs are serving URLLC users, given that RED beats the remaining schemes for low load, but its performance degrades for high load regimes.

CHAPTER 7

CONCLUSION AND PERSPECTIVES

Contents

7.1 Concluding Remarks 73
7.2 Perspectives 75

7.1 Concluding Remarks

In this thesis, we focused on studying resource allocation policies for 5G network slicing while considering multi-connectivity. After presenting an overview of mobile networks' evolution and explaining the novel and unprecedented use cases promised by 5G, we described the mobile network's architecture transformations from distributed RAN to network slicing driven networks. Then, we discussed the ramifications of adopting the new service-based vision on business relationships and the respect of the service level agreement binding the 5G ecosystem's actors. Afterward, we summarized the management entities responsible for slice instantiation to then propose the distribution of these entities taking decisions concerning traffic steering and resource allocation over different architecture segments to allocate resources at the RAN level while enabling multi-tenancy.

In the perspective of delivering URLLC services to users, we explore multi-connectivity, defined as the ability to connect to more than one physical interface simultaneously to achieve stringent reliability requirements. For this reason, we examined a wireless system serving URLLC users covered by two base stations. We model this scenario as a system of coupled queues with the same service time, and we study the impact of redundancy and dynamic policy scheduling on minimizing queuing delays. All things considered, we calculate the decay rates and full distribution of the packets' sojourn time numerically, using contour integrals. Then, we estimate the outage probability corresponding to the required delay threshold for URLLC services based on the results mentioned above. Finally, we evaluate and discuss the performance of allocation schemes suitable for URLLC traffic in terms of the system's load while quantifying the maximum achievable load as a function of reserved bandwidth in the system for a referenced target of outage probability.

Given these points, we assess the architecture impact of management entities responsible for resource allocation regarding URLLC services requirements. We focus on smart factory scenarios where eMBB and URLLC slices coexist to deliver seamless industrial automation. We simulate a bandwidth reservation scenario representing the case where there is hard isolation between traffic belonging to different slices. Thereafter, we studied the shared bandwidth scenario, when the same bandwidth is used for both eMBB and URLLC traffic. Last, we substantiate the bandwidth reservation case results by simulating a system with three base stations serving URLLC users.

This whole study provides, if not a complete self-contained framework, promising scheduling policing emerging from queuing theory to treat various problems linked to network slicing and resource allocation for URLLC services. However, this study only scratches the surface of the network slicing resource allocation concept. First of all, the point of view we followed along this report might be considered narrow. We assessed the multi-connectivity scenario using queuing theory models with an FCFS discipline without incorporating other possible 5G enablers such as advanced waveform technologies, massive MIMO, large bandwidth availability empowered by millimeter-wave spectrum, or beamforming, etc. Indeed, 5G promises significant technical enhancements, which can offer a lot of information and opportunities to improve the overall performance across all types of use cases.

Nonetheless, efficient network slicing is vital to deal with the dynamic characteristics in 5G networks and the high traffic variability caused by the combination of continuous, sporadic, and periodic traffic profiles. Hence, we need to address a challenging dynamic network resource allocation problem to carry out network slicing. The primary challenge is that the traffic arrival characteristics and slice resource requirements in terms of computing, processing, memory, and bandwidth demands can be highly dynamic for each slice. With the dynamics and uncertainty that intrinsically characterize wireless network environments, conventional service and resource management approaches that require complete systems knowledge become ineffective or even impertinent. Therefore, conventional queueing theoretic or model-based optimization becomes intractable with the high reliability and stringent bandwidth and latency requirements, high availability, and strict security imposed by beyond 5G networks. Besides, RAN should support flexible RRM capable of integrating dynamic features to cope with unpredictable network conditions alongside standard RRM functions such as interference management, power control, and mobility control.

7.2 Perspectives

Data-driven methodologies, like reinforcement Learning (RL), which allow network entities to learn and build knowledge about the networks to make optimal decisions, have emerged in recent years. As one of the vital machine learning techniques, RL enables the optimization of an agent's decision-making without prior knowledge of the system and environment. The agent learns to perform actions in an environment based on its state, which represents some of its features, by interacting with it and receiving feedback regarding the performed actions. The agent receives a reward or a penalty for taking a good or bad action, respectively. The agent's goal is to maximize its cumulative reward, also referred to as an expected return. Different reinforcement learning methods yield distinct behaviors for the agent to achieve their goal. These solutions have drawn attention to mobile network research due to their proven efficacy in addressing complex multi-domain problems yielding close to optimal results. When making the decision, the agent then adopts a trial-and-error search for possible optimal state-action pairs, referred to as a policy. Different reinforcement learning methods generate distinct behaviors for the agent to achieve their goal. These solutions have

drawn attention in wireless networks research, given their effectiveness in addressing complex multi-domain problems.

Several improvements can be considered for future research directions using RL techniques. For instance, we can replicate the results concerning multi-connectivity presented in this report, relying on the Q-Learning algorithm, which is amongst the most well-known model-free RL algorithms for computing an optimal policy that maximizes the long-term reward. In this case, the actions would be to send URLLC packets to one BS or the other or duplicate them and send them to both BS. The state of this system can be represented as a vector with the following elements: the number of waiting packets in both BSs (or equivalently the traffic load at each BS), the radio conditions of each packet, and the age of information sent to the scheduler. The latter gets a reward if the packet's waiting time is less than a certain threshold (depending on the use case) or a penalty otherwise. We can also design the reward function to favor packet dispatching instead of systematically duplicating packets to mitigate resource underutilization while avoiding SLA violation.

Nevertheless, Q-Learning suffers from slow convergence speed, especially if the problem's state space and action space are large. The algorithm has to store full tables of an immediate Q-value value, which measure the overall expected reward for each state-action pair. The tables can be too considerable to be maintained on mobile devices. The traditional RL methods struggle to address high-dimensional state spaces representing real-world problems due to their high complexity. Deep Reinforcement Learning aims to solve this problem by employing neural networks as function approximators to reduce the complexity of classical RL methods.

APPENDIX A

LARGE DEVIATION THEORY

The study of large deviations is concerned with the quantification of the probabilities of rare events. Estimates of probabilities of rare events turn out to have an exponential form; *i.e.*, these probabilities decrease exponentially fast as a function of the asymptotic parameter [73]. Consider the following to motivate the exponential form of the large deviations estimates. Let $X_1, X_2, ..X_n$ be a sequence of independent, identically distributed (i.i.d.) random variables with a common distribution function $F_X(x) = \mathbb{P}(X \leqslant x), x \in \mathbb{R}$ and finite mean μ. Fix a number $a > \mu$. Now the probability that is clearly decreasing in in a long-term sense, since by the weak law of large numbers

$$\mathbb{P}\Big(\frac{X_1 + X_2 + \cdots + X_n}{n} \geqslant a\Big) \to 0 \quad \text{as} \quad n \to \infty$$

Fix a positive parameter $\theta > 0$. We have

$$\mathbb{P}\Big(\frac{X_1+\cdots+X_n}{n} > a\Big) = \mathbb{P}(e^{\theta(X_1+\cdots+X_n)} > e^{\theta na}) \tag{A.1}$$
$$\leqslant e^{-\theta na}\mathbb{E}[e^{\theta(X_1+\cdots+X_n)}] \tag{A.2}$$
$$= e^{-\theta na}\big(\mathbb{E}[e^{\theta X_1}]\big)^n \tag{A.3}$$
$$= \big(e^{-\theta a}\mathbb{E}[e^{\theta X_1}]\big)^n \tag{A.4}$$

by the Markov inequality and independence.

This bound is meaningful only if the ratio $\mathbb{E}[e^{\theta X_1}]/e^{\theta a}$ is less than unity. We recognize $\mathbb{E}[e^{\theta X_1}]$ as the moment generating function of X_1 and denote it by $M(\theta)$. For the bound to be useful, we need $\mathbb{E}[e^{\theta X_1}]$ to be at least finite. If we could show that this ratio is less than unity, exponentially fast decay of the probability would be established.

Similarly, suppose we want to estimate

$$\mathbb{P}\Big(\frac{X_1+\cdots+X_n}{n} \leqslant a\Big)$$

for some $a < \mu$. Fixing now a negative $\theta < 0$, we obtain

$$\mathbb{P}\Big(\frac{X_1+\cdots+X_n}{n} \leqslant a\Big) = \mathbb{P}(e^{\theta(X_1+\cdots+X_n)} \geqslant e^{\theta na})$$
$$\leqslant \big(e^{-\theta a}M_(\theta)]\big)^n$$

and now we need to find a negative θ such that $M(\theta) < e^{\theta a}$. In particular, we need to focus on θ for which the moment generating function is finite. For this purpose let $\mathcal{D}(M_X) \triangleq \{\theta : M(\theta) < \infty\}$. Namely $\mathcal{D}(M_X)$ is the set of values θ for which the moment generating function is finite. We call $\mathcal{D}$ the domain of M.

We now establish several properties of the moment generating functions.

Proposition 3. *The moment generating function $M_X(\theta)$ of a random variable X satisfies the following properties:*

1. *$M_X(0) = 1$. If $M(\theta) < \infty$ for some $\theta > 0$ then $M(\theta') < \infty$ for all $\theta' \in [0, \theta]$. Similarly, if $M(\theta) < \infty$ for some $\theta < 0$ then $M(\theta') < \infty$ for all $\theta' \in [\theta, 0]$. In particular, the domain $\mathcal{D}(M)$ is an interval containing zero.*

2. *Suppose $(\theta_1, \theta_2) \subset \mathcal{D}(M_X)$. Then $M(\theta)$ as a function of θ is differentiable in θ for every $\theta_0 \in (\theta_1, \theta_2)$, and furthermore,*

$$\frac{d}{d\theta} M_X(\theta)\big|_{\theta=\theta_0} = \mathbb{E}[Xe^{\theta_0 X}] < \infty.$$

   Namely, the order of differentiation and expectation operators can be changed.

Now suppose the *i.i.d.* sequence $X_i, i \geqslant 1$ is such that $0 \in (\theta_1, \theta_2) \subset D(M)$, where M is the moment generating function of X_1. Namely, M is finite in a neighborhood of 0. Let $a > \mu = \mathbb{E}[X_1]$. Applying Proposition 3, let us differentiate this ratio with respect to θ at $\theta = 0$:

$$\frac{d}{d\theta} \frac{M(\theta)}{e^{\theta a}}\Big|_{\theta=0} = \frac{e^{\theta a}\mathbb{E}[X_1 e^{\theta X_1}] - ae^{\theta a}\mathbb{E}[e^{\theta X_1}]}{e^{2\theta a}} = \mu - a < 0$$

Note that $M(\theta)/e^{\theta a} = 1$ when $\theta = 0$. Therefore, for sufficiently small positive θ, the ratio $M(\theta)/e^{\theta a}$ is smaller than unity, and A.4 provides an exponential bound on the tail probability for the average of $X_1, \ldots, X_n$.

Similarly, if $a < \mu$, the ratio $M(\theta)/e^{\theta a} < 1$ for sufficiently small negative θ.

Theorem 1 (Chernoff bound)**.** *Given an i.i.d. sequence $X_1, \ldots, X_n$ suppose the moment generating function $M(\theta)$ is finite in some interval $(\theta_1, \theta_2) \ni 0$. Let $a > \mu = \mathbb{E}[X_1]$. Then there exists $\theta > 0$, such that $M(\theta)/e^{\theta a} < 1$ and*

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} > a\right) \leqslant \left(\frac{M(\theta)}{e^{\theta a}}\right)$$

Similarly, if $a < \mu$, then there exists $\theta < 0$, such that $M(\theta)/e^{\theta a} < 1$ and

$$\mathbb{P}(\frac{X_1 + \cdots + X_n}{n} < a) \leqslant \big(\frac{M(\theta)}{e^{\theta a}}\big)$$

We have some freedom in choosing θ as long as M is finite in order to make the ratio $M(\theta)/e^{\theta a}$ small. So we could try to find θ which minimizes the ratio $M(\theta)/e^{\theta a}$. The conclusion of the large deviations theory is very often that such a minimizing value θ^* exists and is tight. Namely, it provides the correct decay rate. In this case we can say

$$\mathbb{P}(\frac{X_1 + \cdots + X_n}{n} > a) \approx \exp(-I(a, \theta^*)n),$$

where $I(a, \theta^*) = -\log\big(M(\theta^*)/e^{\theta^* a}\big)$.

Theorem 1 gave us a large deviations bound $(M(\theta)/e^{\theta a})^n$ which we rewrite as $e^{-n(\theta a - \log(M(\theta))}$.

Definition 1. *A Legendre transform of a random variable X is the function*

$$I(a) \triangleq \sup_{\theta \in \mathbb{R}}(\theta a - \log(M(\theta))).$$

The function $I(a)$ is also commonly called the rate function in the theory of Large Deviations.

For more details about large deviation theory and their application in queuing theory, refer to [74, 75]

APPENDIX B

INDUSTRY 4.0 SLICING USE CASE ILLUSTRATION

Industry 4.0 is expected to operate on 5G connectivity to significantly improve the production process through large-scale automation and advanced monitoring techniques. In this appendix, we display a specific slicing framework for the "factories of the future" use case. The latter introduces several challenges as it requires the coexistence of a wide range of applications belonging to different families of technical requirements in the same premise. For instance, the multiplication of the RATs necessitates the definition of a mechanism that flexibly adapts the RAN for each use case.

In particular, a smart factory can host various types of machines and connected devices using both human-centric and Machine-to-Machine (M2M) communication schemes in the same vicinity where only a restrained set of RATs and technologies is suitable for each scenario.

B.1 Use case description

A typical factory is shown in Figure B.1. Inside the factory, a set of small cells using frequencies beyond 6 GHz ensure a partial coverage of the factory,

but with high data rates. In addition to these tenant-owned and managed small cells, outdoor macro BSs belonging to the InP and operating on the sub 2 GHz frequencies ensure complete coverage of the area but with low data rates. Small BSs are connected with macro BSs via the *Xn* interface, and this connection is likely to be ensured via high-capacity wired links. Moreover, if the small cells use proprietary or WiFi technology, the Non-3GPP Interworking Function (N3IWF) can be used for combining both accesses as prescribed in 3GPP Technical Specification 24.502 [76]. Note that the factory may be situated in an industrial area comprising other smart factories that belong to other tenants. The macro network of the InP has to share its resources between the different tenants.

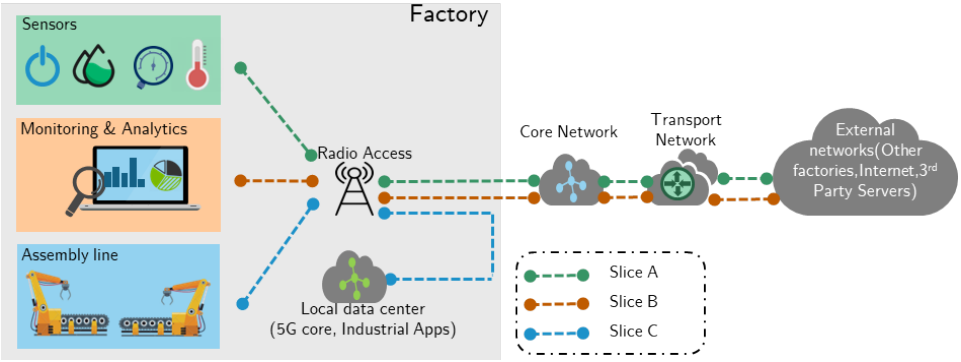


Figure B.1: Industry 4.0 slices from the tenant perspective: the URLLC slice is locally based while the mMTC and eMBB slices reach external networks to the factories premise.

B.1.1 mMTC traffic

The automation of the production process requires a massive number of wireless sensors (temperature, pressure, motion detector, etc.) contending for the same RAN resources and scattered in the factory's premise. The sensors' primary function is to ensure and maintain a continuous production process by detecting anomalies. This particular use case can be distinguished because it produces more uplink than downlink traffic. It should be noted that the generated payload is an aggregation of cyclic and occasional incident-based traffic transmitted using short-size packets and adapted signaling. Equally important, the sensors are not concurrently active, which entails the neces-

sity of predicting the number of arrivals at a given time so that we adequately allocate resources at the RAN. Non-orthogonal multiple access schemes like Sparse Code Multiple Access (SCMA) have been investigated as a possible enabler of the mMTC use case. It allows a large number of devices to connect simultaneously.

B.1.2 eMBB traffic

There is a vast number of eMBB applications that can be deployed for the industry 4.0 use case. These applications meet new KPIs regarding high data rates, wide-area coverage, and high user density. For instance, the data sent by the wireless sensors are treated for real-time monitoring. Furthermore, interactive applications as virtual and augmented reality (VR/AR) and video monitoring can be used in the product design process to simulate, analyze and review the overall production process within the smart factory. The KPIs mentioned above can be achieved by using massive MIMO, large bandwidths of licensed and unlicensed spectrum, combined with advanced modulation schemes.

B.1.3 URLLC traffic

URLLC traffic is generated by the production process and requires extremely low latencies and high reliability. Assembly lines often work on a sequential basis. Thus, any erroneous action stemming from not respecting the E2E latencies or from the use of none-adequate coding schemes may have severe consequences on the factories' performance or may even cause a wide range of severe damages. While the use of new RATs with new frequency bands and subcarrier spacing (15kHz, 30kHz, and 60kHz subcarrier spacing for frequency bands below 6 GHz known and 60 kHz and 120 kHz subcarrier spacing for frequency bands above 24 GHz) will ineluctably reduce the air interface latency, it's crucial to re-design the physical layer to allow faster pre-processing, encoding and decoding times for a smaller frame and flexible structure [70]. Moreover, the latencies associated with the core network will be subdued by deploying the CN functions in a local data center, thus allowing local processing of data and control planes. 3GPP's release 16 [**?**] displays an example of the factory automation technical requirements. For

instance, this use case necessitates reliability of 99.9999%, E2E latency of 2 ms with an air interface latency of 1 ms, short transmissions, a packet size of 32 bytes, and an allowed user equipment speed of 20 m/s. From the RAN resource allocation incentive, we can use the cyclic nature of the URLLC traffic within the factory to perform resource allocation, either by a cyclic reservation of resources for deterministic traffic or by contention-based access with blind replication of packets for a sporadic traffic [77].

B.2 Slice management for the factory

As advocated in [72], a slice corresponds to the combination of a tenant and a service. We have thus to differentiate between the perspectives of the tenant and the InP. Three slices correspond to the three traffic types for the tenant managing the factory and its networking infrastructure. From the InP perspective, there are as many slices as the combination of traffic/tenant couples.

In Figure 4, the slices from the tenant perspective are illustrated. While the URLLC slice is local for allowing low latency, the eMBB and mMTC slices span the whole end-to-end path between the UEs and the application servers that may be located on the Internet, even if they may stay local for some applications. This suggests a partial implementation of the 5G core functions within the factory.

As for radio connectivity, the tenant privileges the usage of its local small cell network for serving its traffic but also uses the macro network of the InP. However, some of its traffic may be steered towards the macro BS in the following cases:

- When the UE is not covered by the small cells, it's connected to the macro BS that forwards its traffic to the Internet or feeds it back to the factory through the Xn interface.
- When the local network is congested, the lowest priority traffic can be steered towards the InP network.
- For some URLLC applications requiring very high reliability, packets

are duplicated and sent to both local and InP BSs to ensure spatial diversity.

This traffic steering policy is defined by the tenant's NSMF and applied by the UEs. The macro BS receives traffic requests from different factories, while the InP's NSSMF takes the scheduling decisions based on the SLAs contracted with separate tenants.

PUBLICATIONS

In what follows, we present the published and submitted papers issued from this doctoral research.

Published papers

- A. Chagdali, S. E. Elayoubi, A. M. Masucci and A. Simonian, "Performance of URLLC Traffic Scheduling Policies with Redundancy," 2020 32nd International Teletraffic Congress (ITC 32), 2020, pp. 55-63, doi: 10.1109/ITC3249928.2020.00015.
- A. Chagdali, S. E. Elayoubi and A. M. Masucci, "Impact of Slice Function Placement on the Performance of URLLC with Redundant Coverage," 2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 2020, pp. 1-6, doi: 10.1109/WiMob50308.2020.9253421.
- A.Chagdali, S.E. Elayoubi, and A. M. Masucci. "Slice Function Placement Impact on the Performance of URLLC with Multi-Connectivity" Computers 10, no. 5: 67. 2021 https://doi.org/10.3390/computers10050067

BIBLIOGRAPHY

[1] Harrison J. Son. Network architecture evolution from 4g to 5g. *NETMANIAS*, December 2015.

[2] 3GPP. Study on new radio access technology: Radio access architecture and interfaces. *3GPP TR 38.801 V14.0.0*, March 2017.

[3] History.com. First speech transmitted by telephone. *AE Television Networks*.

[4] Pekka Lundmark. Thirty years on from the call that transformed how we communicate. July 2021.

[5] Shancang Li, Li Da Xu, and Shanshan Zhao. 5g internet of things: A survey. *Journal of Industrial Information Integration*, 10:1–9, 2018.

[6] Charalampos Kalalas and Jesus Alonso-Zarate. Massive connectivity in 5G and beyond: Technical enablers for the energy and automotive verticals. In *2020 2nd 6G Wireless Summit (6G SUMMIT)*, pages 1–5, 2020.

[7] 5G Americas. New services & applications with 5G ultra-reliable low latency communications. 2018.

[8] Shunliang Zhang. An overview of network slicing for 5g. *IEEE Wireless Communications*, 26(3):111–117, 2019.

[9] C. Chang, N. Nikaein, and T. Spyropoulos. Radio access network resource slicing for flexible service execution. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 668–673, April 2018.

[10] Xenofon Foukas, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K Marina. Network slicing in 5G: Survey and challenges. *IEEE Communications Magazine*, 55(5):94–100, 2017.

[11] Peter Rost, Christian Mannweiler, Diomidis S Michalopoulos, Cinzia Sartori, Vincenzo Sciancalepore, Nishanth Sastry, Oliver Holland, Shreya Tayade, Bin Han, Dario Bega, et al. Network slicing to enable scalability and flexibility in 5G mobile networks. *IEEE Communications magazine*, 55(5):72–79, 2017.

[12] Xuan Zhou, Rongpeng Li, Tao Chen, and Honggang Zhang. Network slicing as a service: enabling enterprises' own software-defined cellular networks. *IEEE Communications Magazine*, 54(7):146–153, 2016.

[13] Jose Ordonez-Lucena, Pablo Ameigeiras, Diego Lopez, Juan J Ramos-Munoz, Javier Lorca, and Jesus Folgueira. Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges. *IEEE Communications Magazine*, 55(5):80–87, 2017.

[14] 3GPP. Summary of rel-15 work items. *3GPP TR 21.915 v15.0.0, Tech. Rep.*, September 2019.

[15] Azad Ravanshid, Peter Rost, Diomidis S. Michalopoulos, Vinh V. Phan, Hajo Bakker, Danish Aziz, Shreya Tayade, Hans D. Schotten, Stan Wong, and Oliver Holland. Multi-connectivity functional architectures in 5g. In *2016 IEEE International Conference on Communications Workshops (ICC)*, pages 187–192, 2016.

[16] Patrick Marsch, Icaro Da Silva, Omer Bulakci, Milos Tesanovic, Salah Eddine El Ayoubi, Thomas Rosowski, Alexandros Kaloxylos, and Mauro Boldi. 5G radio access network architecture: Design guidelines and key considerations. *IEEE Communications Magazine*, 54(11):24–32, 2016.

[17] A. Chagdali, S. E. Elayoubi, and A. M. Masucci. Impact of slice function placement on the performance of URLLC with redundant coverage. In *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 1–6, 2020.

[18] Abdellatif Chagdali, Salah Eddine Elayoubi, Antonia Maria Masucci, and Alain Simonian. Performance of URLLC traffic scheduling policies with redundancy. In *2020 32nd International Teletraffic Congress (ITC 32)*, pages 55–63, 2020.

[19] Abdellatif Chagdali, Salah Eddine Elayoubi, and Antonia Maria Masucci. Slice function placement impact on the performance of URLLC with multi-connectivity. *Computers*, 10(5), 2021.

[20] Nickson M. Karie, Nor Masri Sahri, and Paul Haskell-Dowland. Iot threat detection advances, challenges and future directions. In *2020 Workshop on Emerging Technologies for Security in IoT (ETSecIoT)*, pages 22–29, 2020.

[21] Jorge Navarro-Ortiz, Pablo Romero-Diaz, Sandra Sendra, Pablo Ameigeiras, Juan J. Ramos-Munoz, and Juan M. Lopez-Soler. A survey on 5g usage scenarios and traffic models. *IEEE Communications Surveys Tutorials*, 22(2):905–929, 2020.

[22] NGMN Alliance. Description of network slicing concept. *NGMN 5G P*, 1(1), 2016.

[23] Cisco. Cisco annual internet report (2018–2023). 2020.

[24] Backblaze Patrick Thomas. Defining an exabyte. 2020.

[25] Cheng Liu. *Architectural Evolution and Novel Design of Fiber-Wireless Access Networks*, pages 213–233. 01 2017.

[26] Glauco E. Gonçalves, Guto L. Santos, Leylane Ferreira, Élisson da S. Rocha, Lubnnia M. F. de Souza, André L. C. Moreira, Judith Kelner, and Djamel Sadok. *Flying to the Clouds: The Evolution of the 5G Radio Access Networks*, pages 41–60. Springer International Publishing, Cham, 2020.

[27] Mohammad Asif Habibi, Meysam Nasimi, Bin Han, and Hans D. Schotten. A comprehensive survey of RAN architectures toward 5G mobile communication system. *IEEE Access*, 7:70371–70421, 2019.

[28] Nathan Gomes, Philippe Chanclou, Peter Turnbull, Anthony Magee, and Volker Jungnickel. Fronthaul evolution: From CPRI to ethernet. *Optical Fiber Technology*, 26, 08 2015.

[29] Aleksandra Checko, Henrik Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael Berger, and Lars Dittmann. Cloud RAN for mobile networks—a technology overview. *Communications Surveys Tutorials, IEEE*, 17:405–426, 01 2015.

[30] Hong Ren, Cunhua Pan, Nan Liu, You Xiaohu, Maged Elkashlan, Arumugam Nallanathan, and L. Hanzo. Low-latency **C-RAN**: A next-generation wireless approach. *IEEE Vehicular Technology Magazine*, PP:1–1, 04 2018.

[31] Isiaka Ajewale Alimi, António Luís Teixeira, and Paulo Pereira Monteiro. Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions. *IEEE Communications Surveys Tutorials*, 20(1):708–769, 2018.

[32] Matteo Fiorani, Bjorn Skubic, Jonas Mårtensson, Luca Valcarenghi, Piero Castoldi, Lena Wosinska, and Paolo Monti. On the design of 5G transport networks. *Photonic Network Communications*, 08 2015.

[33] Thomas Pfeiffer. Next generation mobile fronthaul architectures. In *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, pages 1–3, 2015.

[34] Bin Guo, Wei Cao, An Tao, and Dragan Samardzija. LTE/LTE-A signal compression on the CPRI interface. *Bell Labs Technical Journal*, 18(2):117–133, 2013.

[35] Veronica Quintuna Rodriguez, Fabrice Guillemin, Alexandre Ferrieux, and Laurent Thomas. Cloud-RAN functional split for an efficient fronthaul network. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pages 245–250, 2020.

[36] Small Cell Forum. Small cell virtualization: Functional splits and use cases. 2016.

[37] Longsheng Li, Meihua Bi, Haiyun Xin, Yunhao Zhang, Yan Fu, Xin Miao, Ahmed Mohanmmed Mikaeil, and Weisheng Hu. Enabling flexible link capacity for {eCPRI-based fronthaul with load-adaptive quantization resolution. *IEEE Access*, 7:102174–102185, 2019.

[38] M Series. Imt vision–framework and overall objectives of the future development of imt for 2020 and beyond. *Recommendation ITU*, 2083:0, 2015.

[39] Theodore S Rappaport, Shu Sun, Rimma Mayzus, Hang Zhao, Yaniv Azar, Kevin Wang, George N Wong, Jocelyn K Schulz, Mathew Samimi, and Felix Gutierrez. Millimeter wave mobile communications for 5g cellular: It will work! *IEEE access*, 1:335–349, 2013.

[40] Jian Wang, Aixiang Jin, Dai Shi, Lei Wang, Hui Shen, Dan Wu, Liang Hu, Liang Gu, Lei Lu, Yan Chen, et al. Spectral efficiency improvement with 5g technologies: Results from field tests. *IEEE journal on selected areas in communications*, 35(8):1867–1875, 2017.

[41] 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*, 20(4):3098–3130, 2018.

[42] Amna Mughees, Mohammad Tahir, Muhammad Aman Sheikh, and Abdul Ahad. Towards energy efficient 5g networks using machine learning: Taxonomy, research challenges, and future research directions. *IEEE Access*, 8:187498–187522, 2020.

[43] Network Functions Virtualisation NFV. ETSI GS NFV 001 v1. 1.1 (2013-10). 2013.

[44] Diego Kreutz, Fernando M. V. Ramos, Paulo Esteves Veríssimo, Christian Esteve Rothenberg, Siamak Azodolmolky, and Steve Uhlig. Software-defined networking: A comprehensive survey. *Proceedings of the IEEE*, 103(1):14–76, 2015.

[45] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief. A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys Tutorials*, 19(4):2322–2358, 2017.

[46] Phelipe A. de Souza, Abdallah S. Abdallah, Elivelton F. Bueno, and Kleber V. Cardoso. Virtualized radio access networks: Centralization, allocation, and positioning of resources. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6, 2018.

[47] Andres Garcia-Saavedra, Xavier Costa-Perez, Douglas J. Leith, and George Iosifidis. FluidRAN: Optimized vRAN/MEC orchestration. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 2366–2374, 2018.

[48] Dario Sabella, Peter Rost, Yingli Sheng, Emmanouil Pateromichelakis, Umer Salim, Patricia Guitton-Ouhamou, Marco Di Girolamo, and Giovanni Giuliani. RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network. In *2013 Future Network & Mobile Summit*, pages 1–8. IEEE, 2013.

[49] Navid Nikaein, Raymond Knopp, Lionel Gauthier, Eryk Schiller, Torsten Braun, Dominique Pichon, Christian Bonnet, Florian Kaltenberger, and Dominique Nussbaum. Demo: Closer to cloud-RAN: RAN as a service. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, page 193–195, New York, NY, USA, 2015. Association for Computing Machinery.

[50] Navid Nikaein, Mahesh K Marina, Saravana Manickam, Alex Dawson, Raymond Knopp, and Christian Bonnet. Openairinterface: A flexible platform for 5g research. *ACM SIGCOMM Computer Communication Review*, 44(5):33–38, 2014.

[51] Frédéric Pujol, Salah Eddine Elayoubi, Jan Markendahl, and Linda Salahaldin. Mobile telecommunications ecosystem evolutions with 5G. *Communications & Strategies*, (102):109–130, 2016.

[52] View on 5G architecture (version 2.0). *5G PPP Whitepaper*, 2017.

[53] Slawomir Kukliński, Lechoslaw Tomaszewski, Tomasz Osiński, Adlen Ksentini, Pantelis A. Frangoudis, Eleonora Cau, and Marius Corici. A reference architecture for network slicing. In *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pages 217–221, 2018.

[54] Ramon Ferrus, Oriol Sallent, Jordi Perez-Romero, and Ramon Agusti. On 5G radio access network slicing: Radio interface protocol features and configuration. *IEEE Communications Magazine*, 56(5):184–192, 2018.

[55] Konstantinos Samdanis, Xavier Costa-Perez, and Vincenzo Sciancalepore. From network sharing to multi-tenancy: The 5G network slice broker. *IEEE Communications Magazine*, 54(7):32–39, 2016.

[56] Pablo Caballero, Albert Banchs, Gustavo De Veciana, and Xavier Costa-Perez. Network slicing games: Enabling customization in multi-

tenant mobile networks. *IEEE/ACM Trans. Netw.*, 27(2):662–675, April 2019.

[57] Pablo Caballero, Albert Banchs, Gustavo De Veciana, Xavier Costa-Pérez, and Arturo Azcorra. Network slicing for guaranteed rate services: Admission control and resource allocation games. *IEEE Transactions on Wireless Communications*, 17(10):6419–6432, 2018.

[58] 3GPP. Access to the 3GPP 5G core network (5GCN) via non-3GPP access networks (N3AN). *3GPP TS 24.502 V16.2.0, Tech. Spec.*, December 2019.

[59] 3GPP. Study on management and orchestration of network slicing for next generation network. *3GPP TR 28.801 V15.1.0, Tech. Rep.*, January 2018.

[60] L Flatto and HP McKean. Two queues in parallel. *Communications on pure and applied mathematics*, 30(2):255–263, 1977.

[61] Leopold Flatto and S Hahn. Two parallel queues created by arrivals with two demands i. *SIAM Journal on Applied Mathematics*, 44(5):1041–1053, 1984.

[62] Kristen Gardner, Samuel Zbarsky, Sherwin Doroudi, Mor Harchol-Balter, and Esa Hyytia. Reducing latency via redundant requests: Exact analysis. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):347–360, 2015.

[63] Leopold Flatto. Two parallel queues created by arrivals with two demands ii. *SIAM Journal on Applied Mathematics*, 45(5):861–878, 1985.

[64] John F Shortle, James M Thompson, Donald Gross, and Carl M Harris. *Fundamentals of queueing theory*, volume 399. John Wiley & Sons, 2018.

[65] 3GPP. Study on scenarios and requirements for next generation access technologies. *3GPP TR 38.913 v15.0.0, Tech. Rep.*, June 2018.

[66] Jori Selen, Ivo Adan, Stella Kapodistria, and Johan Leeuwaarden. Steady-state analysis of shortest expected delay routing. *Queueing Syst. Theory Appl.*, 84(3–4):309–354, December 2016.

[67] Ivo J. B. F. Adan, Jaap Wessels, and WHM Zijm. Analysis of the asymmetric shortest queue problem. *Queueing Systems*, 8(1):1–58, 1991.

[68] 3GPP. Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC). *3GPP TR 38.824 V16.0.0*, March 2019.

[69] Klaus I Pedersen, Gilberto Berardinelli, Frank Frederiksen, Preben Mogensen, and Agnieszka Szufarska. A flexible 5G frame structure design for frequency-division duplex cases. *IEEE Communications Magazine*, 54(3):53–59, 2016.

[70] Hyoungju Ji, Sunho Park, Jeongho Yeo, Younsun Kim, Juho Lee, and Byonghyo Shim. Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects. *IEEE Wireless Communications*, 25(3):124–130, 2018.

[71] ITU-R. Minimum requirements related to technical performance for IMT-2020 radio interface (s). Technical report, ITU, Tech. Rep., Nov. 2017, 2017.

[72] Salah Eddine Elayoubi, Sana Ben Jemaa, Zwi Altman, and Ana Galindo-Serrano. 5G RAN slicing for verticals: Enablers and challenges. *IEEE Communications Magazine*, 57(1):28–34, 2019.

[73] Massachusetts institute of Technology. Large deviations for i.i.d. random variables. Available at `https://ocw.mit.edu/courses/sloan-school-of-management/15-070j-advanced-stochastic-processes-fall-2013/lecture-notes/MIT15_070JF13_Lec2.pdf`.

[74] A Dembo O Zeitouni and A Dembo. Large deviations techniques and applications. *Applications of Mathematics*, 38, 1998.

[75] Ayalvadi Ganesh and Neil O'Connell. A large deviation principle with queueing applications. *Stochastics and Stochastic Reports*, 73:25–35, 01 2002.

[76] 3GPP. Access to the 3GPP 5G core network (5GCN) via non-3GPP access networks. *3GPP TS 24.502 V15.4.2*, July 2019.

[77] Salah Eddine Elayoubi, Patrick Brown, Matha Deghel, and Ana Galindo-Serrano. Radio resource allocation and retransmission schemes for URLLC over 5G networks. *IEEE Journal on Selected Areas in Communications*, 37(4):896–904, 2019.