

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τηλεπικοινωνιών



Βασιλική Ζαρκαδούλα
9103

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Διαχωρισμός υπηρεσιών δικτύου σε συστήματα Mobile Edge Computing

Επιβλέπων: Γεώργιος Καραγιαννίδης, Καθηγητής

Θεσσαλονίκη, Νοέμβριος 2021

© Βασιλική Ζαρκαδούλα

© Α.Π.Θ.

Διαχωρισμός υπηρεσιών δικτύου σε συστήματα Mobile Edge Computing
Network slicing in Mobile Edge Computing systems

«Η έγκριση της παρούσης Διδακτορικής Διατριβής από το Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης δεν υποδηλώνει αποδοχή των γνώμων του συγγραφέως»

(Ν. 5343/1932, άρθρο 202, παρ. 2)

Aristotle University of Thessaloniki
Department of Electrical Engineering and Computer Science
Division of Telecommunications



Vasiliki Zarkadoula

DIPLOMA THESIS

Network slicing in Mobile Edge Computing systems

Supervisor: Prof. George K. Karagiannidis

Thessaloniki, Greece, November 2021

Περίληψη

Τα τελευταία χρόνια, η ταχεία αύξηση των εφαρμογών ασύρματης σύνδεσης με διάφορες και ανταγωνιστικές απαιτήσεις υπηρεσιών όσον αφορά την απόδοση, το εύρος ζώνης, την αξιοπιστία και την καθυστέρηση έχει καταστήσει τον υπάρχοντα σχεδιασμό δικτύων 4ης γενιάς (4G) «one-size-fits-all», αδύνατο να καλύψει την ανάγκη υποστήριξης διαφορετικών απαιτήσεων ποιότητας υπηρεσιών. Επίσης, η παραδοσιακή Cloud Computing αρχιτεκτονική είναι αδύνατο να ικανοποιήσει τις απαιτήσεις των νέων εφαρμογών σε απόδοση. Σε αυτό το πλαίσιο, οι τεχνολογίες Network Slicing και Mobile Edge Computing (MEC) αποτελούν δύο βασικούς πυλώνες των δικτύων 5ης και μεταγενέστερης γενιάς (5G and beyond), ιδιαίτερα για την ενδυνάμωση υπηρεσιών που χρειάζονται την ελάχιστη δυνατή καθυστέρηση. Η παρούσα εργασία διερευνά τη συνύπαρξη ετερογενών υπηρεσιών σε κοινούς πόρους και υπό ένα Edge Computing σενάριο. Οι υπηρεσίες στα δίκτυα 5ης γενιάς ταξινομούνται σε τρεις κύριες κατηγορίες, συγκεκριμένα, στα Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC) και Massive Machine-Type Communications (mMTC). Στην παρούσα διπλωματική, εξετάζεται η ταυτόχρονη συνύπαρξη χρηστών με πολύ χαμηλές απαιτήσεις σε ενέργεια, οι οποίοι θεωρούμε ότι ανήκουν στην κατηγορία των mMTC και χρηστών με χαμηλές απαιτήσεις σε καθυστέρηση, οι οποίοι θεωρούμε ότι ανήκουν στην κατηγορία των URLLC χρηστών. Το network slicing στοχεύει στην ικανοποίηση των ετερογενών τους απαιτήσεων, ενώ η χρήση ενός edge διακομιστή στην επίτευξη χαμηλών καθυστερήσεων επικοινωνίας. Παράλληλα, μελετώνται τόσο η ορθογώνια όσο και η μη ορθογώνια πολλαπλή πρόσβαση. Ειδικότερα, στην πρώτη περίπτωση, διατυπώνεται ένα κυρτό πρόβλημα βελτιστοποίησης που στοχεύει στην ελαχιστοποίηση του κατωφλίου της καθυστέρησης εκφόρτωσης δεδομένων στον MEC διακομιστή των URLLC χρηστών, ενώ παράλληλα ικανοποιούνται οι ετερογενείς απαιτήσεις και των δύο ειδών συσκευών. Στη συνέχεια, ερευνάται το βέλτιστο “ταίριασμα” των χρηστών και η βέλτιστη κατανομή των πόρων, αξιοποιώντας το σχήμα της μη ορθογώνιας πολλαπλής πρόσβασης. Λαμβάνεται η θεώρηση ότι οι URLLC συσκευές έχουν κατανεμηθεί ορθογώνια σε κανάλια, αλλά μοιράζονται τα κανάλια τους με mMTC συσκευές και διατυπώνεται ένα πρόβλημα βελτιστοποίησης που στοχεύει στην ελαχιστοποίηση της κατανάλωσης ενέργειας για την εκφόρτωση δεδομένων των mMTC χρηστών, ικανοποιώντας ταυτόχρονα τις διαφορετικές απαιτήσεις και των δύο ειδών συσκευών. Το αρχικό πρόβλημα αποσυντίθεται σε δύο υποπροβλήματα και ακολούθως, τα υποπροβλήματα αυτά λύνονται επαναληπτικά έως ότου βρεθεί η βέλτιστη λύση. Τέλος, η αποτελεσματικότητα των προτεινόμενων μεθόδων, επαληθεύεται μέσω αποτελεσμάτων προσομοιώσεων.

Λέξεις - κλειδιά: Ασύρματες Τηλεπικοινωνίες, Βελτιστοποίηση, Μη Ορθογώνια Πολλαπλή Πρόσβαση, Διαμοιρασμός Καναλιών, Κατανομή Πόρων.

Abstract

In recent years, the rapid increase of diverse wireless applications with various and competing service demands in terms of performance, bandwidth, reliability and latency has rendered the existing «one-size-fits-all» 4G network design unable to meet the sheer need of supporting diverse quality of service (QoS) requirements. Furthermore, the traditional Cloud Computing architecture is unable to meet the performance requirements. In this context, Radio Access Network (RAN) slicing and Mobile Edge Computing (MEC) are two key enablers for 5G and beyond, particularly to empower low-latency services. This thesis investigates the coexistence of heterogeneous services on the same radio resource in an Edge Computing scenario. Services in 5G are classified into three main categories, namely, Enhanced Mobile Broadband (eMBB), Massive Machine-Type Communications (mMTC) and Ultra-Reliable Low-Latency Communications (URLLC). In this thesis, the coexistence of users with very low energy requirements, which are considered to belong to the mMTC category, and users with low latency requirements, which are considered to belong to the URLLC category, is examined. Network slicing aims to meet their heterogeneous requirements, while the use of an edge server aims to achieve low communication delays. Meanwhile, both orthogonal (OMA) and non-orthogonal (NOMA) multiple access schemes are examined. In particular, in the first case, a convex optimization problem is formulated, that aims to minimize the MEC offloading delay threshold for URLLC users, while satisfying the heterogeneous requirements of both sets of devices. Then, the optimal pairing and resource allocation for the considered traffic categories are investigated, exploiting the NOMA scheme. URLLC devices are orthogonally assigned to subchannels, but they share their resource blocks with mMTC devices and an optimization problem is formulated, that aims to minimize the energy consumption for the data offloading of mMTC users while satisfying the diverse requirements of both sets of devices. The original problem is decomposed into two sub-problems and these sub-problems are solved iteratively so as to obtain an optimal solution. Finally, the effectiveness of the proposed methods is verified through simulations.

Keywords: 5G, Wireless Communications, Optimization, Network Slicing, Mobile Edge Computing (MEC), Ultra-Reliable Low-Latency Communications (URLLC), Massive Machine-Type Communications (mMTC), Non-orthogonal Multiple Access (NOMA), Pairing, Channel Assignment, Resource Allocation.

Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνονται οι σπουδές μου στον τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών. Αρχικά θα ήθελα να εκφράσω την αμέριστη ευγνωμοσύνη μου στην οικογένεια μου, οι οποίοι πάντα μου προσφέρουν την απόλυτη στήριξή τους σε κάθε μου βήμα. Επιπλέον, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή, κύριο Γεώργιο Καραγιαννίδη για την εμπιστοσύνη που μου έδειξε καθώς και για τις συμβουλές του. Παράλληλα, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Πάυλο Μπουζίνη, για την καθοδήγηση που μου προσέφερε σε όλα τα στάδια εκπόνησης της εργασίας. Τέλος, θα ήθελα να πω ένα μεγάλο ευχαριστώ στις φίλες και στους φίλους που μου χάρισαν πέντε όμορφα φοιτητικά χρόνια.

Βασιλική Ζαρκαδούλα

Περιεχόμενα

Abstract	iii
Ευχαριστίες	v
Περιεχόμενα	vii
Κατάλογος σχημάτων	ix
1 Εισαγωγή	1
1.1 Δίκτυα 5G	1
1.1.1 Enhanced Mobile Broadband	2
1.1.2 Ultra-Reliable Low-Latency Communications	2
1.1.3 Massive Machine-Type Communications	3
1.2 Network Slicing	3
1.2.1 Network Function Virtualization (NFV)	4
1.2.2 Software-Defined Networking (SDN)	4
1.2.3 Συνεργασία SDN - NFV για το Network Slicing	5
1.3 Mobile Edge Computing	5
1.3.1 Από Clouds σε Edges	6
1.3.2 Ορισμός και Χαρακτηριστικά Κλειδιά	6
1.3.3 Σενάρια Χρήσης	7
1.4 Σχετικές Έρευνες και Κίνητρα	8
1.5 Συνεισφορά	10
1.6 Δομή	10
2 Introduction	11
2.1 Towards 5G	11
2.1.1 Enhanced Mobile Broadband	11
2.1.2 Ultra-Reliable Low-Latency Communications	12
2.1.3 Massive Machine Type Communications	12
2.2 Network Slicing	13
2.2.1 Network Function Virtualization (NFV)	13
2.2.2 Software-Defined Networking (SDN)	14
2.2.3 Enabling SDN-NFV for Network Slicing	14
2.3 Mobile Edge Computing	15

2.3.1	From Clouds to Edges	15
2.3.2	Definition and Key Characteristics	15
2.3.3	Use cases	16
2.4	Related Work and Motivation	17
2.5	Contribution	19
2.6	Structure	19
3	URLLC Delay Minimization with FDMA	21
3.1	System Model	21
3.2	Problem Formulation	22
3.3	Simulations and Numerical Results	24
4	mMTC Sum Energy Consumption Minimization with NOMA	29
4.1	Introduction	29
4.2	System Model	29
4.3	Problem Formulation	31
4.4	Proposed Solution	32
4.4.1	Subchannel Assignment for a Fixed Power Allocation	33
4.4.2	Power Allocation for a Fixed Subchannel Assignment	33
4.4.3	Iterative Algorithm Description	37
4.5	Simulations and Numerical Results	38
5	Conclusions	43
A	Hungarian Algorithm Example	45
	Bibliography	47

Κατάλογος σχημάτων

1.1	eMBB, URLLC, mMTC σενάρια χρήσης	3
2.1	eMBB, URLLC, mMTC use cases	13
3.1	Average optimal URLLC latency threshold as a function of mMTC energy consumption threshold for different channel gain coefficients ($E_u = 0.1\text{mJ}, T_m = 0.1\text{sec}$) . . .	25
3.2	Average optimal URLLC latency threshold as a function of URLLC energy consumption threshold for different channel gain coefficients ($E_m = 1\mu\text{J}, T_m = 0.1\text{sec}$)	26
3.3	Average optimal URLLC latency threshold as a function of mMTC latency threshold for different channel gain coefficients ($E_u = 0.1\text{mJ}, E_m = 1\mu\text{J}$)	27
3.4	Average optimal URLLC latency threshold as a function of URLLC (or mMTC) devices for different channel gain coefficients ($E_u = 0.1\text{mJ}, E_m = 1\mu\text{J}, T_m = 0.1\text{sec}$) .	27
3.5	Average optimal URLLC latency threshold as a function of task size of devices for different channel gain coefficients ($E_u = 0.1\text{mJ}, E_m = 1\mu\text{J}, T_m = 0.1\text{sec}$)	28
4.1	NOMA user pairing for MEC	29
4.2	System design of distance setting	31
4.3	Convergence behavior of Algorithm 2 for different numbers of total users ($E_u = 0.1\text{J}, T_u = 0.05\text{sec}, T_m = 1.5\text{sec}$)	39
4.4	Average minimum power consumption of every mMTC and URLLC user in every subchannel for different channel gain coefficients ($E_u = 0.05\text{J}, T_m = 1.5\text{sec}, T_u = 0.03\text{sec}$)	40
4.5	Average minimum sum energy consumption of mMTC users as a function of the resource blocks for different channel gain coefficients ($E_u = 0.05\text{J}, T_u = 0.03\text{sec}, T_m = 1.5\text{sec}$)	41
4.6	Average minimum sum energy consumption of mMTC users as a function of the latency threshold of URLLC users for different channel gain coefficients ($E_u = 0.05\text{J}, T_m = 1.5\text{sec}$)	41
4.7	Average minimum sum energy consumption of mMTC users as a function of the latency threshold of mMTC users for different channel gain coefficients ($E_u = 0.05\text{J}, T_u = 0.03\text{sec}$)	42
4.8	Average minimum sum energy consumption of mMTC users as a function of the task size of users for different channel gain coefficients ($E_u = 0.05\text{J}, T_u = 0.03\text{sec}, T_m = 1.5\text{sec}$)	42

A.1	Hungarian Method - Step 1	45
A.2	Hungarian Method - Step 2	45
A.3	Hungarian Method - Step 3	46
A.4	Hungarian Method - Step 4	46

Κεφάλαιο 1

Εισαγωγή

1.1 Δίκτυα 5G

Τα τελευταία χρόνια, ο αριθμός των συσκευών που διαθέτουν δυνατότητες ασύρματης σύνδεσης έχουν αυξηθεί δραματικά, από κλασικές συσκευές επικοινωνίας, όπως υπολογιστές ή τηλέφωνα, έως οικιακές συσκευές, όπως τηλεοράσεις και ψυγεία. Σε αυτό έχει συμβάλει και η άφιξη νέων υπηρεσιών όπως το διαδίκτυο των πραγμάτων (IoT) και των επικοινωνιών Machine-to-Machine (M2M), όπου κάθε είδους ηλεκτρονικής συσκευής θα έχει δυνατότητα ασύρματης σύνδεσης και επικοινωνίας. Ακόμα περισσότερο, η ασύρματη πρόσβαση έχει γίνει ο κυρίαρχος τρόπος σύνδεσης στο Διαδίκτυο, γεγονός που καθιστά απαραίτητο τα ασύρματα δίκτυα να μπορούν να διαχειριστούν τον υψηλό αυτό όγκο κίνησης και απαιτήσεων.

Αυτές οι αυξανόμενες απαιτήσεις έχουν ληφθεί υπόψη στο σχεδιασμό του νέου δικτύου κινητής τηλεφωνίας πέμπτης γενιάς (5G), το οποίο αναμένεται να υποστηρίξει τις μελλοντικές απαιτήσεις νέων εφαρμογών και υπηρεσιών. Το 5G δεν είναι μόνο μια απλή εξέλιξη της τρέχουσας γενιάς δικτύων, αλλά μπορεί να θεωρηθεί ως μία επανάσταση στην τρέχουσα τεχνολογία πληροφοριών και επικοινωνιών [1]. Σε σύγκριση με τις τρέχουσες τεχνολογίες 4G, το 5G αναμένεται να υποστηρίξει 1000 φορές υψηλότερο όγκο δεδομένων κινητής τηλεφωνίας ανά περιοχή, 10 με 100 φορές πιο γρήγορους ρυθμούς δεδομένων, να μειώσει την καθυστέρηση επικοινωνίας μεταξύ χρηστών 5 φορές και να υποστηρίξει έναν αριθμό συνδεδεμένων συσκευών 10 έως 100 φορές υψηλότερο, χωρίς αύξηση κόστους και κατανάλωσης ενέργειας (10 φορές μεγαλύτερη διάρκεια ζωής μπαταρίας) [2].

Το 5G σχεδιάζεται με γνώμονα την ευελιξία και την επεκτασιμότητα, επιτρέποντας ένα ευρύ φάσμα πιθανών περιπτώσεων χρήσης. Τα χαρακτηριστικά «κλειδιά» του αναγνωρίζονται ως οι πολύ υψηλοί ρυθμοί δεδομένων, η εξαιρετική αξιοπιστία και η χαμηλή καθυστέρηση και η μαζική επικοινωνία μεταξύ συσκευών. Γι' αυτό, έχουν θεσπιστεί τρεις διαφορετικές κατηγορίες υπηρεσιών 5G που καλούνται να καλύψουν αυτές τις ανάγκες, συγκεκριμένα οι enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC) και massive Machine Type Communications (mMTC) [3], [4]. Συνοπτικά, η eMBB υποστηρίζει σταθερές συνδέσεις με υψηλούς ρυθμούς δεδομένων, οι URLLC υποστηρίζουν πολύ χαμηλή καθυστέρηση μετάδοσης από ένα περιορισμένο σύνολο τερματικών και με πολύ υψηλή αξιοπιστία και οι mMTC υποστηρίζουν ένα μεγάλο αριθμό συσκευών Internet of Things (IoT) που είναι μόνο περιστασιακά ενεργές και στέλνουν μικρά πακέτα δεδομένων.

1.1.1 Enhanced Mobile Broadband

Οι eMBB εφαρμογές αποτελούν μια φυσική εξέλιξη των υπάρχοντων 4G δικτύων και παρέχουν ταχύτερους ρυθμούς δεδομένων και συνεπώς καλύτερη εμπειρία χρήστη από τις τρέχουσες ευρυζωνικές υπηρεσίες κινητής τηλεφωνίας. Γι' αυτό και θεωρείται ως η πρώτη από τις τρεις κατηγορίες που θα φέρει τα οφέλη του 5G στο ευρύ κοινό, καθώς μπορεί να προσφέρει υψηλή Quality of Service (QoS) πρόσβαση στο Διαδίκτυο, υπό συνθήκες που μέχρι τώρα μπορεί να ήταν απαιτητικές ή ακόμα και απαγορευτικές¹. Κάποιες από τις βασικές δυνατότητες που προσφέρουν είναι η υψηλότερη χωρητικότητα, η βελτιωμένη συνδεσιμότητα και η υψηλότερη κινητικότητα χρηστών², ενώ εγγυάται μέτρια αξιοπιστία, με ποσοστό σφαλμάτων πακέτων (PER) της τάξεως του 10^{-3} . Συγκεκριμένα, η ευρυζωνική σύνδεση θα προσφέρεται σε πυκνοκατοικημένες περιοχές, τόσο σε εσωτερικούς όσο και σε εξωτερικούς χώρους, όπως κέντρα πόλεων, κτίρια γραφείων ή δημόσιους χώρους, γήπεδα ή συνεδριακά κέντρα και θα είναι διαθέσιμη παντού. Οι κινητές ευρυζωνικές υπηρεσίες θα είναι επίσης διαθέσιμες σε κινούμενα οχήματα όπως αυτοκίνητα, λεωφορεία, τρένα και αεροπλάνα.

1.1.2 Ultra-Reliable Low-Latency Communications

Αποτελούν αναμφισβήτητα τη πιο ελπιδοφόρα και «game-changer», αλλά παράλληλα και τη πιο απαιτητική, προσθήκη στις επερχόμενες δυνατότητες των 5G δικτύων. Οι URLLC εφαρμογές έχουν αυστηρές απαιτήσεις σε αξιοπιστία και σε από άκρη σε άκρη (end-to-end, E2E) καθυστέρηση. Το ιδανικό είναι η επίτευξη της χαμηλότερης δυνατής καθυστέρησης και της υψηλότερης δυνατής αξιοπιστίας, κάτι που πολλές φορές δεν είναι εφικτό λόγω του υψηλού κόστους [5]. Η E2E καθυστέρηση περιλαμβάνει τη καθυστέρηση διάδοσης μέσω ασύρματου δικτύου από τον πομπό στο δέκτη, την καθυστέρηση ουράς και τη καθυστέρηση επεξεργασίας. Αξιοπιστία είναι η δυνατότητα μετάδοσης ενός συγκεκριμένου όγκου κίνησης μέσα σε μία προκαθορισμένη χρονική διάρκεια με μεγάλη πιθανότητα επιτυχίας [6]. Σύμφωνα με το 3GPP, για ένα πακέτο δεδομένων των 32bytes, η αξιοπιστία που προσφέρουν οι URLLC είναι 99.9% και η E2E καθυστέρηση είναι μικρότερη του 1ms [7].

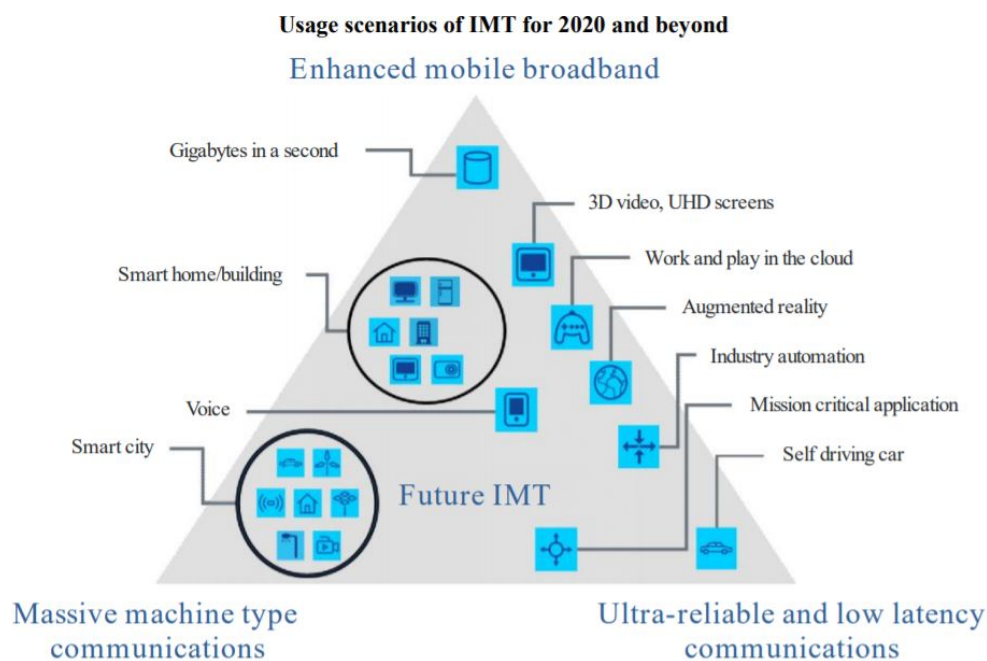
Με την επιτυχία των URLLC, θα ανοίξει ο δρόμος για μία πληθώρα νέων εφαρμογών και θα είναι εφικτή η ψηφιοποίηση ενός ευρέος φάσματος βιομηχανιών. Με τους πολύ χαμηλούς χρόνους, κάτω του 1ms, θα είναι πλέον δυνατή η εξ' αποστάσεως χειρουργική επέμβαση, αφού θα επιτρέπει στους γιατρούς να εξετάζουν το σώμα των ασθενών τους με τη χρήση απτικών ανατροφοδοτήσεων και αισθητήρων σε πραγματικό χρόνο. Πολύ σημαντική θα είναι και η εφαρμογή των URLLC στην βιομηχανία. Ο έλεγχος της βιομηχανίας αυτοματοποιείται με την ανάπτυξη δικτύων σε μονάδες παραγωγής. Συγκεκριμένα, οι URLLC θα έχουν εφαρμογή στην αυτοματοποίηση των εργοστασιακών διαδικασιών και των συστημάτων ισχύος. Επιπλέον, τεχνολογικές αλλαγές θα έρθουν και στον κλάδο μεταφορών, με εφαρμογές όπως τη χρήση drone για την εκτίμηση της πυκνότητας κίνησης σε πραγματικό χρόνο, τα αυτοκινούμενα οχήματα και τον έλεγχο υποσταθμών για τον συγχρονισμό συστημάτων και τη διαχείριση της κυκλοφορίας. Όσον αφορά τον αθλητισμό και τη ψυχαγωγία, το 5G URLLC θα χρησιμοποιείται για μετάδοση ζωντανών εκδηλώσεων, ψυχαγωγία μέσω cloud (Augmented Reality/Virtual Reality), ζωντανές αθλητικές εκδηλώσεις και διαδικτυακά παιχνίδια [6], [8], [9].

¹Source: <https://www.telit.com/blog/5g-embb-use-cases-advantages/>

²Source: <https://5g.co.uk/guides/what-is-enhanced-mobile-broadband-embb/>

1.1.3 Massive Machine-Type Communications

Τα μελλοντικά δίκτυα και τεχνολογίες επικοινωνίας αναμένεται να δώσουν μεγάλη έμφαση στις αποτελεσματικές αλληλεπιδράσεις που αφορούν μηχανές. Σε λίγα χρόνια, ο αριθμός των συνδεδεμένων μηχανών στο διαδίκτυο αναμένεται να ξεπεράσει αυτό των συνδεδεμένων ανθρώπων, φτάνοντας τις δεκάδες δισεκατομμύρια έως το 2025. Επειδή όλα αυτά τα δεδομένα θα μεταδίδονται μέσω ασύρματων δικτύων, τα κυβελωτά δίκτυα που κάποτε σχεδιάστηκαν για να υποστηρίξουν ανθρωποκεντρικές εφαρμογές καλούνται τώρα να επεκταθούν για να υποστηρίξουν έναν τεράστιο αριθμό μηχανών [10]. Οι mMTC λοιπόν, καλούνται να παρέχουν ασύρματη συνδεσιμότητα σε δεκάδες δισεκατομμύρια τύπους συσκευών χαμηλής πολυπλοκότητας και χαμηλής ισχύος και θα αποτελέσουν το θεμέλιο για την επιτυχία του Διαδικτύου των Πραγμάτων (IoT). Πιο συγκεκριμένα, θα δίνεται έμφαση στην κλιμακούμενη συνδεσιμότητα για έναν αυξανόμενο αριθμό συσκευών, στην κάλυψη ευρείας περιοχής και στη βαθιά διείσδυση σε εσωτερικούς χώρους [11]. Επίσης, υποστηρίζουν χαμηλούς ρυθμούς δεδομένων και αποτελούν μια ενεργειακά αποδοτική λύση, καθώς μεγιστοποιούν τη διάρκεια ζωής της μπαταρίας των συνδεδεμένων συσκευών³. Μερικά τυπικά mMTC σενάρια είναι η μακροπρόθεσμη περιβαλλοντική παρατήρηση που περιλαμβάνει περιορισμένη κατανάλωση ενέργειας, οι έξυπνες πόλεις με εκατομμύρια αισθητήρες και τα συνδεδεμένα σπίτια.



M.2083-02

Σχήμα 1.1: eMBB, URLLC, mMTC σενάρια χρήσης

1.2 Network Slicing

Από την προηγούμενη ανάλυση, είναι πλέον προφανές ότι τα δίκτυα 5G στοχεύουν στην υποστήριξη ετερογενών υπηρεσιών. Σε αυτό το πλαίσιο, το network slicing επιτρέπει στους χειριστές των δικτύων να παρέχουν ειδικά εικονικά λογικά δίκτυα με λειτουργίες συγκεκριμένες για μία υπηρεσία

³Source: <https://uk5g.org/discover/5g-topic/massive-machine-type-communications/>

ή για έναν πελάτη, μέσω μιας κοινής υποδομής δικτύου. Κάθε εικονικό δίκτυο (slice), λόγω των λογισμικών που επιτρέπουν την εικονικοποίηση του δικτύου, μπορεί να έχει τη δική του λογική τοπολογία, κανονισμούς ασφαλείας και χαρακτηριστικά απόδοσης, ενώ ταυτόχρονα είναι απομονωμένο από τα άλλα [12]. Αυτό επιτρέπει για παράδειγμα σε ένα slice του δικτύου να παρέχει υπηρεσίες χαμηλής ασφάλειας και χαμηλού εύρους ζώνης (όπως για τις mMTC υπηρεσίες), ενώ ένα άλλο slice μπορεί να παρέχει υπηρεσίες υψηλής ασφάλειας και υψηλής αξιοπιστίας (όπως για τις URLLC υπηρεσίες). Η ιδέα του network slicing φαίνεται να είναι πλήρως υλοποιήσιμη με την άνοδο του Software-Defined Networking (SDN) και του Network Function Virtualization (NFV).

1.2.1 Network Function Virtualization (NFV)

Ο πρωταρχικός στόχος του NFV, όπως υποδηλώνει και το όνομα, είναι η εικονικοποίηση των λειτουργιών του δικτύου οι οποίες μέχρι πρότινος υλοποιούνταν σε hardware. Ειδικότερα, η στρατηγική πίσω από το NFV είναι να αντικαταστήσει ένα σύνολο διαφορετικών στοιχείων που το καθένα έχει και διαφορετικές λειτουργίες, όπως δρομολογητές, μεταγωγείς, τείχη προστασίας και gateways, με εικονικές μηχανές που τρέχουν σε εικονικούς διακομιστές. Οι εικονικές αυτές μηχανές εκτελούν όλες τις λειτουργίες των παραπάνω συσκευών αλλά και πολλών άλλων, είτε με έναν κεντρικό είτε με έναν κατακεντρωμένο τρόπο. Η κεντρική αρχιτεκτονική υποδηλώνει ότι το NFV εκτελείται σε κέντρα δεδομένων, ενώ η κατακεντρωμένη σημαίνει ότι εκτελείται στον εξοπλισμό του χρήστη ή κάπου κοντά του. Επιπλέον, μεγάλο πλεονέκτημα του είναι ότι δεν χρειάζεται εγκατάσταση νέου εξοπλισμού. Ως αποτέλεσμα, η τεχνολογία NFV επιτρέπει στους παρόχους τηλεπικοινωνιακών υπηρεσιών να επεκτείνουν τις δυνατότητες και τις υπηρεσίες δικτύου τους στους καταναλωτές τους με μεγαλύτερη ευελιξία. Επίσης τους επιτρέπει να δημιουργούν νέες υπηρεσίες δικτύου ή να αναβαθμίζουν τις παλαιότερες πιο γρήγορα και με χαμηλότερο κόστος [13],[14],[15].

1.2.2 Software-Defined Networking (SDN)

Η μαζική αύξηση των πολυμέσων, η έκρηξη του cloud computing, ο αντίκτυπος της αυξανόμενης χρήσης κινητών συσκευών και οι συνεχείς πιέσεις των επιχειρήσεων για μείωση του κόστους, ενώ ταυτόχρονα τα έσοδα να παραμένουν σταθερά, αποτελούν μία τεράστια πρόκληση για τα παραδοσιακά επιχειρηματικά μοντέλα. Η στροφή στην τεχνολογία Software-Defined Networking, φαίνεται να είναι η λύση και αυτή που θα φέρει την επανάσταση στο σχεδιασμό και τις λειτουργίες του δικτύου. Το SDN είναι μια αρχιτεκτονική στην οποία ο έλεγχος και η διαχείριση του δικτύου συγκεντρώνονται και διαχωρίζονται από το επίπεδο δεδομένων, επιτρέποντας έτσι στο δίκτυο να είναι προγραμματιζόμενο. Το επίπεδο ελέγχου αποτελείται από έναν ή περισσότερους ελεγκτές, οι οποίοι θεωρούνται ο εγκέφαλος του δικτύου SDN και περιέχουν όλη τη νοημοσύνη του δικτύου. Ωστόσο, όταν πρόκειται για ασφάλεια, επεκτασιμότητα και ευελιξία, ο έξυπνος συγκεντρωτισμός έχει τα δικά του μειονεκτήματα, το οποίο είναι και το θεμελιώδες πρόβλημα του SDN. Ο διαχωρισμός του επιπέδου ελέγχου από το επίπεδο δεδομένων, η χρήση ροών δεδομένων και όχι διευθύνσεων προορισμού για αποφάσεις προώθησης, η μεταφορά της λογικής ελέγχου σε μια εξωτερική οντότητα, ο SDN ελεγκτής και το γεγονός ότι το δίκτυο είναι πλέον προγραμματιζόμενο μέσω εφαρμογών λογισμικού που λειτουργούν στο λειτουργικό σύστημα του δικτύου είναι οι πυλώνες πάνω στους οποίους βασίζεται η παραπάνω τεχνολογία. Οι έννοιες στις οποίες βασίζεται το SDN δεν είναι νέες, αλλά κυρίως αποτέλεσμα προηγούμενων ερευνητικών θεμάτων (π.χ. Openflow). Τα προγραμματιζόμενα

δίκτυα επιταχύνουν την εισαγωγή νέων δυνατοτήτων και υπηρεσιών, ενώ η κεντροποίηση επιτρέπει βελτιστοποίηση της απόδοσης, κλιμακούμενα και ευέλικτα δίκτυα [16],[17].

1.2.3 Συνεργασία SDN - NFV για το Network Slicing

Το SDN και το NFV είναι βασικοί πυλώνες για την επίτευξη της πραγματοποίησης του network slicing. Το SDN παρέχει ευέλικτα και προγραμματιζόμενα δίκτυα 5G. Ο προγραμματισμός διευκολύνει την παροχή υπηρεσιών, τη διαχείριση του δικτύου, την ενσωμάτωση και τα λειτουργικά ζητήματα, ειδικά όταν πρόκειται για την υποστήριξη υπηρεσιών επικοινωνίας. Το Open Networking Foundation (ONF) έχει αναπτύξει ένα πλήρες πρότυπο SDN για την ανάλυση του 5G network slicing. Στην αρχιτεκτονική ONF, κάθε «περιεχόμενο πελάτη» SDN αντιπροσωπεύει μια πιθανή φέτα δικτύου (network slice). Με τον όρο «περιεχόμενο του πελάτη» νοείται το εννοιολογικό συστατικό ενός διακομιστή που αντιπροσωπεύει όλες τις πληροφορίες για ένα δεδομένο πελάτη και είναι υπεύθυνο για τη συμμετοχή σε ενεργές λειτουργίες διαχείρισης-ελέγχου διακομιστή-πελάτη. Ο ελεγκτής SDN διαχειρίζεται τα slices χρησιμοποιώντας ένα σύνολο κανόνων ή πολιτικών και απλοποιεί τη δημιουργία περιεχομένων πελάτη και διακομιστή, καθώς και την εγκατάσταση των σχετικών πολιτικών τους. Ειδικότερα, συντηρεί το περιεχόμενο πελάτη ενός slice δικτύου και έτσι μπορεί να διαχειριστεί δυναμικά τα slices ομαδοποιώντας αυτά που έχουν το ίδιο περιεχόμενο. Ο ελεγκτής SDN επίσης «κυβερνά» τα slices του και ενορχηστρώνει τη διαχείριση πόρων στο περιεχόμενο του διακομιστή. Επιπλέον, η δυνατότητα προγραμματισμού που παρέχεται από την τοπολογία SDN επιτρέπει σε τρίτα μέρη να ελέγχουν τους καταναμημένους πόρους, όπως τη δικτύωση και τους πόρους cloud μέσω ανοιχτών API. Αυτό, με τη σειρά του, καθιστά δυνατή την προσανατολισμένη και κατά παραγγελία παροχή υπηρεσιών και την ελαστικότητα των πόρων στα εικονικά και προγραμματιζόμενα 5G δίκτυα.

Παρόλο που η αρχιτεκτονική SDN παρέχει μια πλήρη εικόνα των χαρακτηριστικών του επιπέδου ελέγχου που επιτρέπουν το slicing, δεν διαθέτει δυνατότητες που είναι κρίσιμες για την αποτελεσματική διαχείριση του κύκλου ζωής των slices του δικτύου και των πόρων τους. Από αυτή την άποψη, η αρχιτεκτονική NFV είναι κατάλληλη για αυτόν τον ρόλο, επειδή διαχειρίζεται τους πόρους υποδομής και ενορχηστρώνει την κατανομή τους για την υλοποίηση Virtual Network Functions (VNFs) και υπηρεσιών δικτύου. Μια σωστή συνεργασία μεταξύ SDN και NFV είναι απαραίτητη έτσι ώστε οι πάροχοι να επωφεληθούν από τις δυνατότητες διαχείρισης και ενορχήστρωσης του NFV. Ωστόσο, η ενσωμάτωση των τεχνολογιών αυτών σε ένα κοινό πλαίσιο αναφοράς είναι ένα δύσκολο έργο. Το ETSI πρότεινε ένα προκαταρκτικό πλαίσιο για την ενσωμάτωση του SDN στην αρχιτεκτονική αναφοράς NFV, αλλά η επισκόπηση του παραλείπεται [18], [19], [20], [21].

1.3 Mobile Edge Computing

Προκειμένου να συμβαδίσουν με την ταχεία πρόοδο της τεχνολογίας, οι κατασκευαστές σχεδιάζουν συσκευές που παράγουν και καταναλώνουν έναν αυξανόμενο όγκο δεδομένων. Ο αριθμός των συσκευών είναι τεράστιος και κάθε μία έχει το δικό της σύνολο απαιτήσεων. Οι συσκευές μπορεί επίσης να βρεθούν σε διάφορες τοποθεσίες σε όλο τον κόσμο και ορισμένες έχουν πολύ χαμηλή ανοχή σε καθυστερήσεις, όπως για παράδειγμα οι συσκευές που επιτελούν εφαρμογές σε πραγματικό χρόνο. Αυτές οι συσκευές είναι μέρος μεγαλύτερων συστημάτων τα οποία απαιτούν υπολογιστική ισχύ και αποθηκευτικό χώρο για την επεξεργασία και την αποθήκευση δεδομένων. Δεκάδες δισεκατομμύρια

τέτοιες συσκευές «Edge» αναμένεται να αναπτυχθούν στο εγγύς μέλλον, καθώς οι CPU ταχύτητες αυξάνονται εκθετικά σύμφωνα με τον νόμο του Moore [22].

1.3.1 Από Clouds σε Edges

Μέχρι ενός σημείου, η επεξεργασία των δεδομένων μπορεί να εκτελεστεί στη συσκευή. Για πολλές εφαρμογές αυτό όμως δεν είναι αρκετό και απαιτούνται περισσότεροι υπολογιστικοί πόροι. Το Cloud Computing θεωρείται μία αρχική επιλογή εύρεσης αυτών των πόρων. Το Cloud Computing είναι η διάθεση υπολογιστικών πόρων μέσω διαδικτύου, ιδιαίτερα χώρου για αποθήκευση δεδομένων και υπολογιστικής ισχύς, από κεντρικά συστήματα που βρίσκονται απομακρυσμένα από τον τελικό χρήστη και διαχειρίζονται από παρόχους υπηρεσιών cloud ⁴. Ωστόσο, έχει έναν εγγενή περιορισμό, που είναι η μεγάλη απόσταση διάδοσης μεταξύ του τελικού χρήστη και του απομακρυσμένου cloud, με αποτέλεσμα την εξαιρετικά μεγάλη χρονική καθυστέρηση. Ως συνέπεια, είναι ανεπαρκές για ένα ευρύ φάσμα νέων εφαρμογών κινητών συσκευών που απαιτούν χαμηλές καθυστερήσεις επικοινωνίας [23]. Σε αυτό το πλαίσιο, μπορεί κανείς να επωφεληθεί από τη συσσωρευμένη τεράστια ποσότητα αδρανούς επεξεργαστικής ισχύος και αποθηκευτικού χώρου στα άκρα του δικτύου έτσι ώστε να εκτελέσει υπολογιστικά ευαίσθητες και κρίσιμες σε καθυστέρηση εργασίες στις φορητές συσκευές. Αυτό το πρωτόκολλο είναι γνωστό ως Mobile Edge Computing (MEC).

1.3.2 Ορισμός και Χαρακτηριστικά Κλειδιά

Το Mobile Edge Computing αναγνωρίζεται από τον Ευρωπαϊκό Ερευνητικό Φορέα 5G PPP (5G Infrastructure Public Private Partnership) ως μία από τις βασικές αναδυόμενες τεχνολογίες για δίκτυα 5G (μαζί με το NFV και το SDN) ⁵. Σύμφωνα με το Ευρωπαϊκό Ινστιτούτο Τηλεπικοινωνιακών Προτύπων (ETSI), «το *Mobile Edge Computing* παρέχει δυνατότητες *Cloud Computing* εντός του Δικτύου Ασύρματης Πρόσβασης (*Radio Access Network, RAN*) και σε κοντινή απόσταση από τις κινητές συσκευές. Ο στόχος τους είναι η μείωση της καθυστέρησης, η διασφάλιση της αποδοτικής λειτουργίας του δικτύου, η παροχή υπηρεσιών και η βελτιωμένη εμπειρία χρήστη » [24]. Το MEC λειτουργεί στην άκρη του RAN, παρέχοντας υπηρεσίες υπολογισμών, αποθήκευσης και δικτύωσης. Οι διακομιστές MEC λειτουργούν σε μια γενική υπολογιστική πλατφόρμα και εγκαθίστανται απευθείας στους σταθμούς βάσεις, επιτρέποντας την εκτέλεση εφαρμογών κοντά στους τελικούς χρήστες. Αναλυτικότερα, όπως αναφέρεται από το ETSI, το MEC χαρακτηρίζεται από:

- Την πλήρωση αυστηρών απαιτήσεων χαμηλής καθυστέρησης των χρηστών, καθώς οι υπηρεσίες Edge εκτελούνται κοντά στις τελικές συσκευές
- Απομόνωση από το υπόλοιπο δίκτυο, ενώ προσφέρεται πρόσβαση σε τοπικούς πόρους (πολύ σημαντικό για σενάρια επικοινωνίας μεταξύ μηχανών)
- Εγγύτητα, η οποία προσφέρει το πλεονέκτημα της ανάλυσης και της εφαρμογής μεγάλων δεδομένων

⁴Πηγή: <https://www.ibm.com/cloud/learn/cloud-computing>

⁵Πηγή: <https://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf>

- Επίγνωση τοποθεσίας, καθώς ο MEC διακομιστής λαμβάνει πληροφορίες από Edge συσκευές στο τοπικό δίκτυο πρόσβασης και βρίσκει τη θέση τους
- Πληροφορίες περιβάλλοντος δικτύου, καθώς εφαρμογές και υπηρεσίες μπορούν να χρησιμοποιούν δεδομένα δικτύου σε πραγματικό χρόνο για την παροχή υπηρεσιών που σχετίζονται με το περιβάλλον τους

1.3.3 Σενάρια Χρήσης

Υπάρχουν πολλές υπηρεσίες που θα μπορούσαν να επωφεληθούν από ένα κατανεμημένο σύννεφο κοντά στους πελάτες. Αυτή η ενότητα συζητά μερικά από τα πιο σημαντικά σενάρια εφαρμογής.

Επαυξημένη Πραγματικότητα

Η επαυξημένη πραγματικότητα (Augmented Reality, AR) είναι ο συνδυασμός μιας προβολής του πραγματικού περιβάλλοντος και πρόσθετων αισθητηριακών εισόδων που δημιουργούνται από υπολογιστή, όπως ήχος, βίντεο, γραφικά ή δεδομένα GPS. Η τεχνολογία επαυξημένης πραγματικότητας απαιτεί οι συσκευές να αναλύουν οπτικά δεδομένα (π.χ. κάμερα της συσκευής) και να ενσωματώνουν προκαθορισμένα οπτικά στοιχεία σε πραγματικό χρόνο. Η χρήση μιας MEC πλατφόρμας μειώνει τον χρόνο επιστροφής και αυξάνει την απόδοση, παρέχοντας υψηλότερη ποιότητα εμπειρίας χρήστη. Μπορεί να χρησιμοποιηθεί για την παροχή καταναλωτικών ή επιχειρηματικών προτάσεων, όπως τουριστικές πληροφορίες, πληροφορίες αθλητικών εκδηλώσεων, διαφημίσεις και ούτω καθεξής [24].

Streaming

Η ανάλυση του video streaming έχει υψηλή υπολογιστική πολυπλοκότητα και καταναλώνει μεγάλο εύρος ζώνης και επομένως δεν είναι πρακτικό να πραγματοποιείται η ανάλυσή τους ούτε στη συσκευή ή ούτε στο κεντρικό cloud. Εκτελώντας ανάλυση βίντεο κοντά στις Edge συσκευές, το σύστημα μπορεί όχι μόνο να επωφεληθεί από τη μειωμένη καθυστέρηση, αλλά και να αποφύγει το πρόβλημα της συμφόρησης του δικτύου που προκαλείται από το βίντεο.

Αυτόνομα οχήματα

Αυτά τα οχήματα πρέπει να συλλέγουν και να αναλύουν τεράστιους όγκους δεδομένων σχετικά με το περιβάλλον τους, τις κατευθύνσεις και τις καιρικές συνθήκες, καθώς και να επικοινωνούν με άλλα οχήματα. Πρέπει επίσης να αλληλεπιδρούν με τοπικά δημοτικά δίκτυα και να παρέχουν δεδομένα στους κατασκευαστές προκειμένου να ειδοποιούνται για τη χρήση και τις ανάγκες συντήρησης. Ο MEC σχεδιασμός επιτρέπει στα αυτόνομα οχήματα να συλλέγουν, να επεξεργάζονται και να διανέμουν δεδομένα σε πραγματικό χρόνο, να μεταδίδουν προειδοποιήσεις κινδύνου και να καλύπτουν ευαίσθητες σε καθυστέρηση επικοινωνίες, με μια καθυστέρηση 20 ms από άκρο σε άκρο, λαμβάνοντας και αξιολογώντας σήματα από κοντινά οχήματα και αισθητήρες στο δρόμο [23].

Διαδίκτυο των πραγμάτων

Η ετερογένεια των συσκευών που εκτελούν διάφορα πρωτόκολλα είναι ένα σημαντικό χαρακτηριστικό του Διαδικτύου των Πραγμάτων (Internet of Things, IoT), επομένως η διαχείρισή τους θα πρέπει να αντιμετωπίζεται από μια πύλη χαμηλής καθυστέρησης, όπως ο MEC διακομιστής. Η MEC αρχι-

τεκτονική μπορεί να χρησιμοποιηθεί για την επεξεργασία και τη συγκέντρωση των μικρών πακέτων δεδομένων που αποστέλλονται από υπηρεσίες IoT, προτού αυτά φτάσουν στο κεντρικό δίκτυο. Καθώς ο αριθμός των συνδέσεων IoT αυξάνεται, η χρήση μιας MEC λογικής θα είναι ζωτικής σημασίας για την επεκτασιμότητα και τη διάρκεια ζωής των συσκευών IoT που λειτουργούν με μπαταρία. Η διάρκεια ζωής της συσκευής μπορεί να παραταθεί λόγω του ότι ο μικρότερος χρόνος επικοινωνίας μεταξύ της συσκευής και του διακομιστή εφαρμογών μειώνει την εξάντληση της μπαταρίας [25].

Υγεία

Ορισμένα σύνολα δεδομένων που σχετίζονται με την υγειονομική περίθαλψη είναι τόσο ζωτικής σημασίας που οποιαδήποτε καθυστέρηση στην ανάλυση και κατανόησή τους δεν είναι ανεκτή. Επομένως, η λύση του MEC φαίνεται να είναι πολύ πρακτική. Για την προστασία του απορρήτου των δεδομένων, ένας διακομιστής MEC κοντά στο νοσοκομείο μπορεί να επεξεργάζεται δεδομένα τοπικά και παρέχει στους επαγγελματίες ειδοποιήσεις σε πραγματικό χρόνο για απρόσμενες συμπεριφορές ασθενών, καθώς και πίνακες ελέγχου ασθενών 360 μοιρών για πλήρη ορατότητα. Η υπολογιστική στα άκρα μπορεί να αναγνωρίσει ποια δεδομένα δεν είναι ζωτικής σημασίας, όπως οι μετρήσεις ενός κανονικού καρδιακού ρυθμού, ενώ ταυτόχρονα εντοπίζει, επεξεργάζεται και αντιδρά σε κρίσιμα δεδομένα, ειδοποιώντας τους κλινικούς γιατρούς να ενεργήσουν το συντομότερο δυνατό ⁶.

1.4 Σχετικές Έρευνες και Κίνητρα

Τα τελευταία χρόνια, έχουν διεξαχθεί αρκετές μελέτες σχετικά με την κατανομή πόρων και το network slicing για τους eMBB, mMTC και URLLC χρήστες. Στο [26], ο Popovski και άλλοι μελετούν και συγκρίνουν την κοινή χρήση πόρων μέσω ορθογωνίας (Orthogonal Multiple Access, OMA) και μη ορθογωνίας (Non Orthogonal Multiple Access, NOMA) ασύρματης πρόσβασης για επικοινωνίες άνω ζεύξης, όταν μια ομάδα συσκευών eMBB, mMTC και URLLC συνδέονται σε έναν μόνο σταθμό βάσης. Συγκεκριμένα, οι συγγραφείς ερεύνησαν δύο σενάρια, δηλαδή το slicing για την ικανοποίηση των eMBB και URLLC συσκευών και το slicing για την ικανοποίηση των eMBB και mMTC συσκευών. Ωστόσο, δεν εξέτασαν το slicing για συσκευές URLLC και mMTC.

Η συνύπαρξη των υπηρεσιών eMBB και URLLC έχει ερευνηθεί σε [27], [28] και σε πολλές άλλες έρευνες. Οι συγγραφείς στο [27] εξετάζουν τη μη ορθογωνία συνύπαρξη eMBB και URLLC σε τοπολογίες fog-radio, όπου τα δεδομένα των URLLC επεξεργάζονται στην άκρη του δικτύου, ενώ των eMBB διαχειρίζονται κεντρικά σε ένα υπολογιστικό σύννεφο. Το πρόβλημα της κοινής συνύπαρξης eMBB/URLLC συσκευών μελετάται στο [28], διερευνώντας εναλλακτικά μοντέλα για την απώλεια της ταχύτητας μετάδοσης των eMBB που προκαλείται από την ταυτόχρονη μετάδοση των URLLC. Επίσης, η συνύπαρξη eMBB και mMTC διερευνάται σε [29], [30] και αρκετές άλλες μελέτες. Ο στόχος της έρευνας [29] είναι να διερευνήσει πολυάριθμες στρατηγικές κατανομής πόρων σε sliced δίκτυο τυχαίας πρόσβασης, να εντοπίσει τον αντίκτυπό τους στην απόδοση στα slices του δικτύου και να προτείνει μια ιδέα Τυχαίας Πρόσβασης Καναλιού (Random Access Channel, RACH). Υπό ένα σενάριο ενός σταθμού βάσης πολλαπλών κεραιών, οι συγγραφείς στο [30] εξετάζουν την απόδοση της άνω ζεύξης των δύο υπηρεσιών eMBB και mMTC, η οποία μετριέται με βάση τους εφικτούς

⁶Πηγή: https://stlpartners.com/edge_computing/10-edge-computing-use-case-examples/

ρυθμούς μετάδοσης δεδομένων και τον αριθμό των συνδεδεμένων συσκευών, για συγκεκριμένες απαιτήσεις αξιοπιστίας.

Επιπλέον, οι ερευνητικές δραστηριότητες έχουν αντιμετωπίσει και τη συνύπαρξη υπηρεσιών mMTC και URLLC σε μια κοινή φυσική υποδομή και από την προοπτική της κατανομής πόρων σε ένα δίκτυο ασύρματης πρόσβασης, αλλά είναι περιορισμένες. Στο [31], η συνύπαρξή τους εξετάζεται υιοθετώντας ένα σχήμα μη ορθογώνιας πολλαπλής πρόσβασης με πολυπλεξία στην ισχύ (power-domain NOMA) μέσα σε ένα μπλοκ κοινόχρηστων πόρων, όπου κάθε υπο-φέρον μοιράζεται τόσο από συσκευές που είναι ευαίσθητες σε καθυστέρηση (URLLC) όσο και από συσκευές που έχουν ανοχή σε καθυστερήσεις επικοινωνίας (mMTC). Ωστόσο, μόνο δύο συσκευές μπορούν να αντιστοιχιστούν στον ίδιο υπο-φέρον, περιορίζοντας έτσι την αύξηση του αριθμού των συσκευών που έχουν πρόσβαση. Έτσι, προκειμένου να αυξηθεί ο αριθμός των συνδέσεων, οι συγγραφείς στο [32] επιτρέπουν σε πολλές συσκευές mMTC και URLLC να έχουν πρόσβαση στον ίδιο υπο-φέρον. Η εργασία [33] προτείνει μια matching game λύση για την άνω ζεύξη στα συστήματα IoT στενής ζώνης (Narrowband IoT, NB-IoT) με σκοπό την αύξηση της χωρητικότητας και του ρυθμού μετάδοσης δεδομένων, προσφέροντας συνδεσιμότητα για έναν τεράστιο αριθμό υπηρεσιών MTC (mMTC και URLLC).

Έχει επίσης διερευνηθεί η κατανομή των πόρων σε διαφορετικά Mobile Edge Computing σενάρια. Οι περισσότερες έρευνες εστιάζουν στο συνδυασμό MEC και NOMA και χρησιμοποιούν το OMA για σύγκριση. Μέχρι στιγμής, οι πτυχές του NOMA-MEC που έχουν ερευνηθεί σχετίζονται κυρίως με τη ελαχιστοποίηση της ενέργειας και της καθυστέρησης. Για παράδειγμα, οι συγγραφείς στο [34] προτείνουν έναν επαναληπτικό αλγόριθμο αναζήτησης για τη μείωση της μέγιστης καθυστέρησης αποστολής πακέτων των χρηστών, βελτιστοποιώντας τον χωρισμό των πακέτων τους και την ισχύ μετάδοσής τους. Με την από κοινού βελτιστοποίηση των φορτίων των χρηστών και του χρόνου μετάδοσης με NOMA, η εργασία [35] στοχεύει στη μείωση της συνολικής καθυστέρησης του χρήστη. Οι συγγραφείς στο [36] σχεδίασαν ένα πρόβλημα ελαχιστοποίησης καθυστέρησης για την εκφόρτωση δεδομένων ως ένα fractional programming πρόβλημα και συνέκριναν το καθαρό NOMA (pure NOMA) με το υβριδικό NOMA (hybrid NOMA) και το OMA. Η μελέτη [37] εξετάζει τις παρεμβολές εντός μίας κυψέλης και προτείνει μια τεχνική βελτιστοποίησης για την κατανομή των υπολογιστικών και επικοινωνιακών πόρων, προκειμένου να ελαχιστοποιηθεί η κατανάλωση ενέργειας των χρηστών που στέλνουν δεδομένα σε έναν MEC διακομιστή. Στο [38], η συνολική ενέργεια του συστήματος μειώθηκε βελτιστοποιώντας την κατανομή ισχύος, την κατανομή του χρόνου μετάδοσης και τη ποσότητα των δεδομένων που εκφορτώνονται στα άκρα του δικτύου. Αφού βρήκαν τις βέλτιστες μαθηματικές σχέσεις για την ισχύ, οι συγγραφείς χρησιμοποίησαν έναν αλγόριθμο successive convex approximation για να αποκτήσουν τη βέλτιστη κατανομή του χρόνου και του ποσού εκφόρτωσης δεδομένων. Οι συγγραφείς του [39] παρέχουν έναν ευρετικό αλγόριθμο, ο οποίος περιλαμβάνει την κατανομή των μπλοκ πόρων, της ισχύος μετάδοσης και των υπολογιστικών πόρων, για την ελαχιστοποίηση της συνολικής ενέργειας σε ένα σύστημα NOMA-MEC.

Ακολουθώντας την λογική όπως στην έρευνα του Popovski [26], σε αυτήν την εργασία εστιάζουμε στο slicing των πόρων επικοινωνίας του Δικτύου Ασύρματης Πρόσβασης. Επίσης, εμπνευσμένοι από τις πρόσφατες έρευνες που μελετούν σενάρια για την άνω ζεύξη συσκευών MTC, σε αυτήν την εργασία μελετάμε την απόδοση του ορθογώνιου και μη ορθογώνιου RAN network slicing σε μία κυψέλη, όπου πολλαπλές συσκευές mMTC και URLLC επικοινωνούν στην άνω ζεύξη με μία κεραία

ενός σταθμού βάσης. Σε αντίθεση με προηγούμενες μελέτες, σε αυτήν την εργασία αντιμετωπίζουμε το πρόβλημα της κατανομής πόρων σε δίκτυα 5G χρησιμοποιώντας τόσο το network slicing όσο και το Mobile Edge Computing.

1.5 Συνεισφορά

Οι κύριες συνεισφορές περιγράφονται παρακάτω.

- Εξετάζεται ένα σύνολο χρηστών URLLC, αλλά δίνεται έμφαση μόνο στις απαιτήσεις τους για χαμηλή καθυστέρηση και δεν λαμβάνονται υπόψη οι απαιτήσεις τους για αξιοπιστία. Αντίστοιχα, εξετάζεται ένα σύνολο συσκευών mMTC και δίνεται έμφαση στην ικανοποίηση των απαιτήσεων χαμηλής κατανάλωσης ενέργειας, αντί να αντιμετωπίζεται το ποσοστό άφιξής τους όπως στις περισσότερες μελέτες.
- Αρχικά, μελετάται ένα σενάριο άνω ζεύξης όπου πολλοί χρήστες URLLC και mMTC εκφορτώνουν τις εργασίες τους σε έναν διακομιστή MEC, χρησιμοποιώντας το σχήμα ορθογώνιας πολλαπλής πρόσβασης συχνότητας (FDMA). Συγκεκριμένα, διατυπώνεται ένα κυρτό πρόβλημα βελτιστοποίησης που στοχεύει στην ελαχιστοποίηση του κατωφλίου καθυστέρησης των χρηστών URLLC.
- Επιπλέον, μελετάται ένα μη ορθογώνιο σχήμα πολλαπλής πρόσβασης που εξυπηρετεί χρήστες URLLC και mMTC σε ένα σενάριο MEC. Πιο συγκεκριμένα, διατυπώνεται και λύνεται ένα νέο πρόβλημα βελτιστοποίησης. Ο στόχος της βελτιστοποίησης είναι να ελαχιστοποιηθεί η κατανάλωση ενέργειας των mMTC χρηστών ενώ παράλληλα ικανοποιούνται οι απαιτήσεις σε ενέργεια και καθυστέρηση όλων των χρηστών.
- Μέσω προσομοιώσεων αξιολογείται η απόδοση της ορθογώνιας και της μη ορθογώνιας πρόσβασης και επιβεβαιώνεται η ανώτερη απόδοσή τους σε σύγκριση με επιλεγμένα σημεία αναφοράς.

1.6 Δομή

Η δομή αυτής της διπλωματικής εργασίας οργανώνεται ως εξής: Στο Κεφάλαιο 3 εξετάζεται η περίπτωση της ελαχιστοποίησης του κατωφλίου καθυστέρησης των URLLC συσκευών χρησιμοποιώντας το σχήμα της ορθογώνιας πολλαπλής πρόσβασης. Παρέχεται το μοντέλο του συστήματος, η διατύπωση του προβλήματος βελτιστοποίησης και η μαθηματική απόδειξη της κυρτότητας. Παρουσιάζονται και συζητούνται τα αποτελέσματα της προσομοίωσης. Το Κεφάλαιο 4 εξετάζει την περίπτωση της ελαχιστοποίησης της ενέργειας εκφόρτωσης δεδομένων στον MEC διακομιστή χρησιμοποιώντας το σχήμα της μη ορθογώνιας πολλαπλής πρόσβασης. Παρουσιάζεται το μοντέλο του συστήματος και διατυπώνεται ένα πρόβλημα βελτιστοποίησης. Στη συνέχεια περιγράφεται η προτεινόμενη λύση σε αυτό το πρόβλημα και παρουσιάζονται τα αποτελέσματα της προσομοίωσης προκειμένου να αξιολογηθούν και να συγκριθούν οι προτεινόμενες μέθοδοι. Τέλος, τα συμπεράσματα παρουσιάζονται στο Κεφάλαιο 5 και στη συνέχεια, στο παράρτημα παρέχεται ένα παράδειγμα εύρεσης του βέλτιστου “ταιριάσματος” με τη χρήση του αλγόριθμου Hungarian .

Chapter 2

Introduction

2.1 Towards 5G

In recent years, the number of devices with wireless capabilities has risen dramatically, ranging from traditional communication devices, such as computers and telephones, to home appliances, such as televisions and refrigerators. The introduction of new services like Internet of Things (IoT) and Machine-to-Machine (M2M) communications, where any type of electronic device will be able to connect and communicate wirelessly, has also contributed to this. Even more, wireless access has become the dominant way of connecting to the Internet, necessitating the need for wireless networks to be able to handle such high volumes of traffic and demand.

These evolving needs have been included into the design of the new fifth-generation mobile networks (5G), which are expected to meet the needs of new apps and services in the future. 5G infrastructure is more than just an evolution of previous network generations; it is a revolution in the field of information and communication technology (ICT) [1]. Compared to current 4G technologies, 5G is expected to provide 1000 times higher mobile data volume per area, 10 to 100 times higher data rates, to reduce 5 times the end-to-end latency and to support 10 to 100 times higher number of connected devices, without increasing the cost and the power consumption (10 times longer battery life) [2].

5G is built with scalability and flexibility in mind, enabling for a wide range of applications. High data rates, high reliability and low latency, and mass connectivity between devices are considered its key characteristics. Therefore, three different categories of 5G services have been developed to suit these demands, namely enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low-Latency Communications (URLLC), and massive Machine Type Communications (mMTC) [3], [4]. In general, eMBB enables fixed links with high data rates, URLLC offers very low transmission delays and very high reliability for a limited set of terminals, and mMTC supports a large number of Internet of Things (IoT) devices that are only occasionally active and deliver small packets of data.

2.1.1 Enhanced Mobile Broadband

eMBB applications are a natural evolution of existing 4G networks, offering higher data rates and thus a better user experience than current mobile broadband services. As a result, they are

regarded as the first of the three categories to bring the benefits of 5G to the general public, since they can provide high quality of service (QoS) Internet access under conditions that were previously difficult or even prohibitive¹. Some of their key features are higher capacity, improved connectivity and higher user mobility², while guaranteeing moderate reliability, with an order of 10^{-3} packet error rate (PER). In detail, broadband access will be available everywhere, even in densely populated areas, both indoors and outdoors, such as city centers, office buildings or public spaces, stadiums, or conference centers. Furthermore, mobile broadband services will be available in moving vehicles such as cars, buses, trains, and planes.

2.1.2 Ultra-Reliable Low-Latency Communications

They are, without a doubt, the most promising and game-changing, but also the most demanding addition to the upcoming capabilities of 5G networks. URLLC applications must meet stringent reliability and end-to-end (E2E) latency requirements. The lowest possible latency and the highest possible reliability is the optimal solution. However, due to the high cost, it is unlikely to be an effective one [5]. E2E latency includes transmission delay over the wireless network from the transmitter to the receiver, queuing delay, processing/computing delay and retransmissions. The capability of sending a specific amount of data in a predetermined time duration with a high probability of success is defined as reliability.

The success of URLLC will bring a plethora of new applications and they will digitize a wide range of industries. The targeted 1ms latency (or lower) is critical in the use of haptic feedback and real-time sensors to allow doctors to examine patients' bodies from a remote operating room. URLLC's deployment in industry is also vital. Industry control is automated by establishing networks in factories. Factory, process, and power system automation are examples of common industrial automation use cases that require URLLC. Technology advancements will impact the transportation sector too, including applications such as the usage of drones for real-time traffic estimation, motor vehicles, and the control of sub-stations for system synchronization and traffic management. In terms of sports and entertainment, 5G URLLC will be used for live reporting of events, live sports events, online gaming and cloud-based entertainment (VR/AR) [6], [8], [9].

2.1.3 Massive Machine Type Communications

Future networks and communication technologies are predicted to place a strong emphasis on machine-to-machine interactions. The number of connected machines to the internet is predicted to surpass the number of connected people by 2025. Because all of this data will be transferred over wireless networks, cellular networks that were formerly meant to support human-centered applications will now be required to expand in order to accommodate a large number of machines [10]. Thus, mMTCs will be the ones to provide wireless connectivity to tens of billions of low-complexity, low-power devices, and will be the foundation for the success of the Internet of Things (IoT). More specifically, mMTC solutions must enable scalable connectivity for a growing number of devices, wide-area coverage and deep indoor penetration [11]. They also support low rates and are an energy efficient solution as they maximize lifespan of the battery of the connected

¹Source: <https://www.telit.com/blog/5g-emb-emb-use-cases-advantages/>

²Source: <https://5g.co.uk/guides/what-is-enhanced-mobile-broadband-emb/>

devices ³. Some typical mMTC scenarios are long-term environmental observation involving limited energy consumption, smart cities with millions of sensors and connected homes.

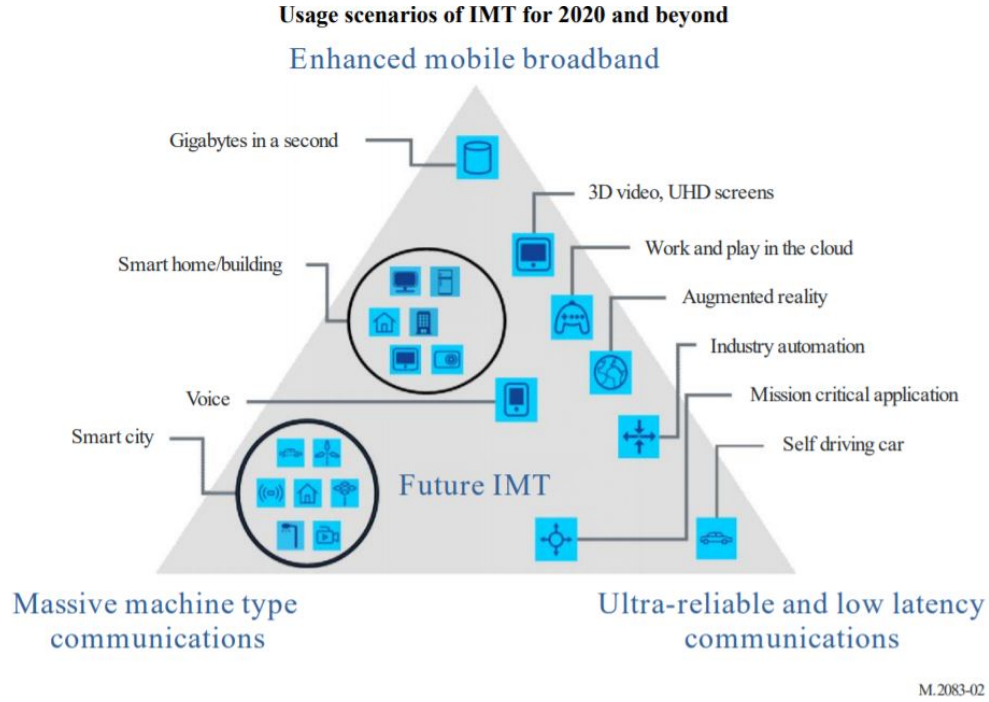


Figure 2.1: eMBB, URLLC, mMTC use cases

2.2 Network Slicing

From the previous analysis, it is now evident that 5G networks aim to support heterogeneous services. In this context, network slicing enables operators to provide dedicated virtual logical networks with functionality specific to the service or customer over a shared network infrastructure. Each virtual network (slice) can have its own logical topology, security regulations, and performance characteristics while being isolated from the others, as a result of the network virtualization software implementations [12]. This allows one network slice to provide low-security and low-bandwidth services (such as for mMTC), while another slice can provide high-security, high-reliability services (such as for URLLC). The concept has been more fully realized with the rise of Software-Defined Networking (SDN) and Network Function Virtualization (NFV).

2.2.1 Network Function Virtualization (NFV)

The primal goal of NFV, as the name suggests, is the virtualization of network functions which were formerly implemented by hardware. In particular, the strategy behind NFV is to replace a set of different elements that each functions differently, such as routers, switches, firewalls and gateways, with virtual machines (VMs) on virtualized servers that will perform all of the above's devices functions and of other's devices too, either in a centralized or in a distributed manner. The centralized architecture suggests that NFV is executed in data centers, while the distributed

³Source: <https://uk5g.org/discover/5g-topic/massive-machine-type-communications/>

one implies that it is executed at the UE or somewhere in proximity to the UE. The fact that there is no need of installation of new equipment is of great importance. Thus, NFV technology allows telecommunications service providers to expand their network capabilities and services to their consumers with greater flexibility. It also allows them to establish new network services or upgrade older ones more quickly and at a lower cost.

2.2.2 Software-Defined Networking (SDN)

The massive growth in multimedia content, the explosion of cloud computing, the impact of rising mobile usage, and ongoing business pressures to cut costs while revenues remain flat are all causing havoc on traditional business models. Turning to SDN technology in order to revolutionize network design and operations seems to be the solution. SDN is an architecture in which control and network management are centralized and separated from the data plane, thus allowing the network to be programmable. The control plane is composed of one or more controllers, which are regarded to be the brain of the SDN network and contain all of the network's intelligence. However, when it comes to security, scalability, and flexibility, intelligent centralization has its own downsides, which is the fundamental issue with SDN. The separation of the control plane from the data plane, the use of data streams rather than destination addresses for forwarding decisions, the transfer of control logic to an external entity, the SDN controller, and the fact that the network is now programmable through software applications running on the network operating system are the pillars upon which the above technology is built. The concepts underlying SDN are not new, but rather the result of previous research topics (e.g. Openflow). Programmable networks accelerate the introduction of new features and services, while centralization enables performance optimization, scalable, and flexible networks [16],[17].

2.2.3 Enabling SDN-NFV for Network Slicing

SDN and NFV are key enablers to achieve the realization of network slices. SDN provides flexible and programmable 5G networks. Programmability facilitates service delivery, network management, integration and operational issues, especially when it comes to supporting communication services. The Open Networking Foundation (ONF) has developed a thorough SDN paradigm for the 5G network slicing analysis. In the ONF architecture, every SDN client context represents a potential slice. Client context is a server's conceptual component that represents all information about a specific client and is responsible for active server-client management and control actions. The SDN controller manages network slices using a set of rules or policies and it simplifies the creation of both server and client contexts as well as the installation of their associated policies. The SDN controller, in particular, maintains a network slice client context and thus it can dynamically manage network slices by grouping slices belonging to the same context. The SDN controller governs its slices and orchestrates resource management on the server context. Furthermore, the programmability provided by SDN topology allows third parties to control the allocated slice resources, such as networking and cloud resources via open APIs. This, in turn, enables on-demand service-oriented customisation and resource elasticity on 5G softwarized and virtualized networks.

Although the SDN architecture provides a full view of the control plane features that enable slicing, it lacks capabilities that are critical for efficiently managing the lifecycle of network slices and their constituent resources. In this regard, the NFV architecture is suited for this role because it manages infrastructure resources and orchestrates their allocation to realize Virtual Network Functions (VNFs) and network services. A proper collaboration between SDN and NFV is necessary to benefit from the management and orchestration capabilities of NFV. However, incorporating SDN and NFV technologies into a common reference framework is a challenging task. ETSI has proposed a preliminary framework for integrating SDN into the reference NFV architecture, but its overview is omitted [18],[19],[20],[21].

2.3 Mobile Edge Computing

In order to keep up with the rapid advancement of technology, manufacturers are designing devices that produce and consume an increasing amount of data. The number of devices is enormous, and each one has its own set of requirements based on the role it serves. Devices may be also found in a variety of locations across the globe and some have a poor delay tolerance as they need to perform, for example, real-time applications. These devices are components of larger systems that require computational power and storage space in order to process and store data. Tens of billions of such Edge devices are expected to be deployed in the near future, with CPU speeds increasing exponentially in accordance with Moore's Law [22].

2.3.1 From Clouds to Edges

Up to a point, data processing can be executed on the device, but for many applications this is not enough and more computing resources are needed. Cloud Computing is considered as an initial option for finding these resources. Cloud computing is the on-demand access, via the Internet, to computer system resources—especially data storage (cloud storage) and computing power—hosted at a remote data center and managed by a cloud services provider⁴. However, Cloud Computing has an intrinsic limitation, which is the large propagation distance from the end user to the remote cloud center, resulting in extremely long latency for mobile applications. As a result, it is insufficient for a wide range of new mobile applications that require low latency [23]. In this context, sufficient capacity to perform computationally sensitive and latency critical tasks of mobile devices can be provided by accumulating the massive amount of idle processing power and storage space distributed at the network edges. This protocol is known as Mobile Edge Computing.

2.3.2 Definition and Key Characteristics

MEC is recognized by the European 5G PPP (5G Infrastructure Public Private Partnership) research body as one of the key emerging technologies for 5G networks (together with Network Functions Virtualization (NFV) and Software-Defined Networking (SDN))⁵. According to the European Telecommunications Standards Institute (ETSI), «*Mobile Edge Computing provides*

⁴Source: <https://www.ibm.com/cloud/learn/cloud-computing>

⁵Source: <https://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf>

an IT service environment and cloud-computing capabilities at the edge of the mobile network, within the Radio Access Network (RAN) and in close proximity to mobile subscribers. The aim is to reduce latency, ensure highly efficient network operation and service delivery, and offer an improved user experience» [24]. MEC operates at the edge of the mobile RAN, providing computing, storage, and networking services. MEC servers run on a generic computing platform and are installed directly at the BSs, allowing applications to be executed close to end users. In more detail, as stated by ETSI, MEC is characterised by:

- Meeting strict low-latency requirements of users, as Edge services run close to end devices
- Isolation from the rest of the network, while having access to local resources (very important for machine-to-machine (M2M) scenarios)
- Proximity, which offers the advantage of analyzing and implementing big data
- Location awareness, as MEC server receives information from edge devices within the local access network and finds the location of the devices
- Network Context Information, as applications and services can use real-time network data to provide context-related services

2.3.3 Use cases

There are numerous services that could benefit from being hosted in a distributed cloud close to customers. This section discusses some of the most important application scenarios.

Augmented Reality

Augmented reality (AR) is the combination of a view of the real-world surroundings and additional computer-generated sensory input such as audio, video, graphics, or GPS data. AR technology necessitates that devices analyse visual data (e.g. device's camera) and incorporate pre-rendered visual elements in real time. Using a MEC platform reduces round-trip time and increases throughput, providing the highest quality experience. It can be used to provide consumer or business propositions such as tourist information, sporting event information, advertisements, and so on [24].

Streaming

Video stream analysis has high computation compexity and consumes a great amount of bandwidth, if these processing tasks are sent to the central cloud. Therefore, it is not practical to perform their analysis neither to the device, nor to the central cloud. By executing video analysis close to the edge devices, the system may not only profit from reduced latency, but also avoid the problem of network congestion caused by video.

Autonomous Vehicles

These vehicles need to collect and analyze massive volumes of data regarding their surroundings, directions, and weather conditions, as well as communicate with other vehicles. They also have to interact with local municipal networks and provide data back to manufacturers in order to track usage and maintenance alerts. Edge computing design enables autonomous vehicles to collect, process, and distribute data in real time, broadcast hazard warnings and latency-sensitive communications within a 20ms end-to-end delay by receiving and evaluating signals from nearby vehicles and roadside sensors [23].

Internet of Things

The heterogeneity of devices running various protocols is an important feature of IoT, thus their management should be handled by a low-latency aggregation point (gateway) such as the MEC server. MEC can be used to process and aggregate the short packets sent by IoT services before they reach the core network. As the number of IoT connections rises, this will be crucial for scalability and battery-powered IoT devices. The device's life can be extended due to fact that the shorter communication time between the device and the application server reduces battery draining [25].

Healthcare

Some healthcare related data sets are so vital that any delay in analyzing and comprehending them cannot be tolerated. Therefore, the edge computing solution appears to be very practical. To protect data privacy, an edge server on the hospital site might process data locally and provides practitioners with real-time notifications of unexpected patient trends or behaviors, as well as with 360-degree view patient dashboards for complete visibility. Edge computing can recognize which data points aren't crucial, such as normal heart rate measurements, while also identifying, processing, and reacting to critical data points, notifying clinicians to act on them as soon as possible ⁶.

2.4 Related Work and Motivation

In recent years, several studies have been conducted regarding the resource allocation and network slicing for eMBB, mMTC and URLLC users. In [26], Popovski et al. study and compare the orthogonal (OMA) and non-orthogonal (NOMA) radio access network (RAN) resource sharing in uplink communications, when a group of eMBB, URLLC, and mMTC devices are connected to a single base station. In particular, the authors investigated two scenarios, namely slicing between eMBB and URLLC and slicing between eMBB and mMTC. However, authors did not consider slicing between URLLC and mMTC devices.

The coexistence of eMBB and URLLC services has been researched in [27],[28] and in many other papers. The authors in [27] examine the non-orthogonal coexistence of eMBB and URLLC in fog-radio topologies, where URLLC traffic is processed at the edge, while eMBB traffic is

⁶Source: https://stlpartners.com/edge_computing/10-edge-computing-use-case-examples/

handled centrally at a cloud. The joint eMBB/URLLC scheduling problem is studied in [28], by exploring alternative models for the eMBB rate loss caused by URLLC superposition/puncturing. The coexistence of eMBB and mMTC is also investigated in [29], [30] and other several studies. The goal of research [29] is to deal with numerous resource allocation strategies in a sliced Random Access Network, identify their impact on slice performance and propose a Random Access Channel (RACH) concept. In a situation with a multi-antenna Base Station, authors in [30] examine the uplink performance of both services, which is measured in terms of achievable data rates and the number of connected devices for the given reliability requirements.

Futhermore, research activities have addressed the coexistence of mMTC and URLLC services in a common physical infrastructure from the RAN resource allocation perspective, but they are limited. In [31], their coexistence is considered by adopting power-domain NOMA within a shared resource block, where each sub-carrier can be shared by both delay sensitive (URLLC) and delay tolerant (mMTC) devices. However, only two devices can be assigned to the same subcarrier, thus limiting the increase in the number of access. Therefore, in order to increase the number of connections, the authors in [32] allow numerous mMTC and URLLC devices to access the same sub-carrier. The work of [33] proposes a matching game solution for the uplink in NB-IoT systems based on the NOMA scheme to increase capacity and data rate by offering additional connectivity for a huge number of MTC (mMTC and URLLC) devices.

The resource allocation in different Mobile Edge Computing scenarios has also been investigated. Most of the works focus on the combination of MEC and NOMA and use OMA for comparison. So far, the aspects of NOMA-MEC that have been researched are mostly related to energy conservation and delay minimization. For instance, authors in [34] propose a bisection search iterative algorithm to reduce the maximum task latency among users by optimizing their tasks partition ratios and offloading transmit power. By jointly optimizing the users' offloaded workloads and the NOMA transmission-time, the work of [35] aims to reduce the overall user latency. As a form of fractional programming, the authors in [36] designed a delay minimization problem for NOMA-MEC data offloading and compared pure NOMA to hybrid NOMA and OMA. The study [37] considers intra-cell interferences and proposes a joint optimization technique for allocating computing and communication resources in order to minimize the energy consumption of MEC users. In [38], the total system energy was reduced by optimizing power allocation, transmission time allocation and offloaded task portions. After obtaining the optimal powers, authors used a successive convex approximation algorithm to obtain the optimal time allocation and offloading task partitions. Authors in [39] provide a heuristic algorithm for the total energy minimization which includes the allocation of transmission resource blocks, transmission power and computational resources in a NOMA assisted MEC system.

Following the premiss as in the work of Popovski et al. [26], in this thesis we focus on the slicing of RAN communication resources for wireless access. Also, inspired on the recent works that study MTC uplink scenarios with receive diversity, in this work we study the performance of orthogonal and non-orthogonal network slicing in a single-cell scenario where multiple mMTC and URLLC devices communicate in the uplink with a single-antenna BS. In contrast to previous studies, we address the problem of resource allocation in 5G networks using both network slicing and Mobile Edge Computing.

2.5 Contribution

The main contributions are described below.

- A set of URLLC users is considered, but only their low latency requirements are addressed and their reliability requirements are not taken under consideration. Accordingly, a set of mMTC devices is considered and the focus is on satisfying their low energy consumption demands, rather than addressing their traffic arrival rate, as in most studies.
- Initially, an uplink scenario is examined, where multiple URLLC and mMTC users offload their tasks, using the orthogonal multiple access scheme in frequency domain (FDMA), to a MEC server. In particular, a convex optimization problem is formulated, that aims to minimize the latency threshold of URLLC users.
- Furthermore, a non-orthogonal multiple access scheme is examined, where URLLC and mMTC users are served in a MEC scenario. More specifically, an optimization problem is formulated and solved. The goal of the optimization is to minimize the energy consumption of mMTC users under the energy and latency requirements of all users.
- Through simulations, the performance of the orthogonal and non-orthogonal slicing is evaluated and their superior performance is confirmed when compared to selected benchmarks.

2.6 Structure

The structure of this thesis is organized as follows: In Chapter 3, the case of URLLC delay minimization using an orthogonal multiple access scheme is considered. The system model, the problem formulation and the mathematical proof of convexity are provided. Simulation results are presented and discussed. Chapter 4 examines the case of mMTC energy minimization using a non-orthogonal multiple access scheme. The system model and the formulated optimization problem are presented. The proposed solution to this problem is then described and simulation results are presented in order to evaluate and compare the proposed methods. Finally, the conclusions are presented in Chapter 5 and after that, in the appendices, a simple step-by-step example of the pairing process by using the Hungarian Algorithm is provided.

Chapter 3

URLLC Delay Minimization with FDMA

3.1 System Model

We consider a MEC offloading scenario in which a set of $\mathcal{K} = \{1, \dots, K\}$ mobile device users with heterogeneous requirements have a computationally intensive task to complete. Specifically, we divide the set of mobiles into two subsets. We assume that the first subset consists of $\mathcal{U} = \{1, \dots, U\}$ URLLC users, so the second subset will consist of $\mathcal{M} = \{U + 1, \dots, K\}$ mMTC users. URLLC devices need to achieve the lowest possible latency, i.e. the time needed for data transmission to the edge cloud, processing and downloading of results to each URLLC mobile device, under a tolerable energy consumption constraint. On the other hand, mMTC devices need to consume the lowest possible energy, subject to a latency constraint, not as strict as for URLLC users. It should be noted that the downloading phase is expected to be insignificant when compared to the offloading and processing phases and is hence ignored.

There is also a single-antenna base station (BS) that is the gateway of an edge cloud, through which the mobile device users can offload the computation. Let F denote the computation capacity of the edge cloud measured by the number of CPU cycles per unit time and F_k the computation capacity of the edge cloud allocated to the k_{th} user, so that

$$\sum_{k=1}^K F_k \leq F \quad (3.1)$$

The overall system bandwidth B is partitioned into K orthogonal subchannels satisfying

$$\sum_{k=1}^K B_k \leq B \quad (3.2)$$

where B_k is the bandwidth of the k_{th} subchannel. Each subchannel is assigned to at most one user, either to a URLLC one or to an mMTC one. Each user k offloads L_k -bit input data. We assume that the transmit power of each device k is denoted as p_k and it follows,

$$0 < p_k \leq p_{\max,k}, \quad \forall k \in \mathcal{K} \quad (3.3)$$

Then the achievable rate (in bits/sec) follows:

$$r_k = B_k \log_2 \left(1 + \frac{p_k g_k}{B_k N_o} \right), \quad \forall k \in \mathcal{K} \quad (3.4)$$

where N_o is the power spectral density of the additive white gaussian noise (AWGN) and $g_k = |h_k|^2 d_k^{-a}$ is the channel gain, where d_k denotes the distance between the k_{th} user and the BS, a is the path loss factor and h_k is the exponential channel gain corresponding to Rayleigh fading. The time required to offload the data to the cloud is given by

$$t_{\text{off},k} = \frac{L_k}{r_k}, \quad \forall k \in \mathcal{K} \quad (3.5)$$

and the computing time at the cloud is

$$t_{\text{comp},k} = \frac{C_k L_k}{F_k}, \quad \forall k \in \mathcal{K} \quad (3.6)$$

where C_k denotes the the number of CPU cycles required to process a single bit of input data of the k – *th* mobile. So the delay, in total, of mobile k is defined as:

$$\begin{aligned} t_k &= t_{\text{off},k} + t_{\text{comp},k} \\ &= \frac{L_k}{r_k} + \frac{C_k L_k}{F_k} \\ &= \frac{L_k}{B_k \log_2 \left(1 + \frac{p_k g_k}{B_k N_o} \right)} + \frac{C_k L_k}{F_k}, \quad \forall k \in \mathcal{K} \end{aligned} \quad (3.7)$$

These delays must satisfy the latency requirements $t_u \leq T_u, \forall u \in \mathcal{U}$ and $t_m \leq T_m, \forall m \in \mathcal{M}$, where T_u and T_m are the maximum acceptable latencies for URLLC and mMTC users respectively and it holds that $T_u \ll T_m$. Finally, the energy consumed for the offload can be expressed as follows:

$$E_{\text{off},k} = p_k t_{\text{off},k} = \frac{p_k L_k}{B_k \log_2 \left(1 + \frac{p_k g_k}{B_k N_o} \right)}, \quad \forall k \in \mathcal{K} \quad (3.8)$$

It is noted that the energy consumption for the computation at the MEC server is omitted. We let E_u and E_m be the energy consumption thresholds for URLLC and mMTC users respectively. Then, it follows that $E_{\text{off},u} \leq E_u, \forall u \in \mathcal{U}$ and $E_{\text{off},m} \leq E_m, \forall m \in \mathcal{M}$, where $E_m \ll E_u$.

3.2 Problem Formulation

In this section, we focus on the minimization of the latency threshold T_u of URLLC users, under constraints of energy consumption, latency, transmission power, bandwidth and CPU capacity for both subsets. In more detail, as mentioned in the above section, we limit the total delay of every user k by a threshold T_k , which is stricter for URLLC devices and the energy consumption by a threshold E_k , which is stricter for mMTC devices. The allocated CPU capacity and bandwidth and the transmission power of each device are constrained as described in (3.1), (3.2) and (3.3) respectively. Our goal is to find the optimal power, bandwidth and cloud computation capacity allocation in order to obtain the optimal minimum URLLC latency threshold. Thus, the corresponding problem can be formulated as

$$\begin{aligned}
& \min_{\{\mathbf{T}_u, \mathbf{B}_k, \mathbf{F}_k, \mathbf{p}_k\}} T_u \\
& \text{s.t.} \\
& C1: t_u \leq T_u \Rightarrow \frac{L_u}{B_u \log_2 \left(1 + \frac{p_u g_u}{B_u N_o}\right)} + \frac{C_u L_u}{F_u} \leq T_u, \quad t_u \geq 0, \quad \forall u \in \mathcal{U} \\
& C2: t_m \leq T_m \Rightarrow \frac{L_m}{B_m \log_2 \left(1 + \frac{p_m g_m}{B_m N_o}\right)} + \frac{C_m L_m}{F_m} \leq T_m, \quad t_m \geq 0, \quad \forall m \in \mathcal{M} \\
& C3: E_{\text{off},u} \leq E_u \Rightarrow \frac{p_u L_u}{B_u \log_2 \left(1 + \frac{p_u g_u}{B_u N_o}\right)} \leq E_u, \quad E_{\text{off},u} \geq 0, \quad \forall u \in \mathcal{U} \quad (\mathbf{P1}) \\
& C4: E_{\text{off},m} \leq E_m \Rightarrow \frac{p_m L_m}{B_m \log_2 \left(1 + \frac{p_m g_m}{B_m N_o}\right)} \leq E_m, \quad E_{\text{off},m} \geq 0, \quad \forall m \in \mathcal{M} \\
& C5: p_k \leq p_{\max,k}, \quad p_k > 0, \quad \forall k \in \mathcal{K} \\
& C6: \sum_{k=1}^K B_k \leq B, \quad B_k \geq 0, \quad \forall k \in \mathcal{K} \\
& C7: \sum_{k=1}^K F_k \leq F, \quad F_k \geq 0, \quad \forall k \in \mathcal{K}
\end{aligned}$$

where $\mathbf{T}_u, \mathbf{B}_k, \mathbf{F}_k, \mathbf{p}_k$ are vectors related with the latency threshold of URLLC, allocated bandwidth, allocated CPU frequency at each user and the transmission power of each user respectively. The first two constraints restrict the overall delay and the next two the offloading energy consumption of URLLC and mMTC devices. Constraint $C5$ defines the upper and lower bounds of the transmission power of each user, while $C6$ and $C7$ guarantee that the allocated bandwidth and CPU frequency at each mobile user is limited to the maximum bandwidth and CPU capacity respectively.

Lemma 1. *Problem P1 is jointly convex with respect to T_u, B_k, F_k, p_k .*

Proof. It is easy to verify that the objective function and the constraints $C5, C6, C7$ are convex, as they are linear functions w.r.t. T_u, p_k, B_k and F_k respectively. The function

$$g_1(B, p) = B \log_2 \left(1 + \frac{pg}{BN_o}\right) \quad (3.9)$$

is concave because its Hessian matrix is negative semi-definite, i.e. its eigenvalues are non-positive:

$$\lambda_{1,2} = \left\{ -\frac{(p^2 + B^2)g^2}{\ln 2 B(BN_o + gp)^2}, 0 \right\} \leq 0 \quad (3.10)$$

The function $1/g_1$ is convex, when g_1 is a positive concave function, thus the first terms of constraints $C1$ and $C2$ are convex. Moreover, the function

$$g_2(F) = \frac{CL}{F} \quad (3.11)$$

is also convex, as its second derivative is positive:

$$g_2''(F) = \frac{2CL}{F^3} > 0 \quad (3.12)$$

It is straightforward that the function $g_3(T) = -T$ is linear, thus it is both convex and concave. Therefore, constraints $C1$ and $C2$ are convex as sums of convex functions. In order to prove the convexity of $C3$ and $C4$ we can define a function

$$g_4(B, p) = pL - EB \log_2 \left(1 + \frac{pg}{BN_o} \right) \quad (3.13)$$

such that constraints $C3$ and $C4$ can be written in the form: $g_4(B_u, p_u) \leq 0$ and $g_4(B_m, p_m) \leq 0$. The first term of g_4 is a linear function w.r.t. p , while the second term is a term of the form: $-Eg_1(p, B)$, where E is a positive constant and g_1 is concave, as proven above. When a function g_1 is concave, then $-g_1$ is convex. Consequently, the second term is also convex and finally, constraints $C3$ and $C4$ are convex, as a sum of convex functions and the proof is completed. \square

Since the optimization problem is convex, we conclude that there is a unique optimal latency threshold T_u^* . The nature of the problem, however, prevents the development of closed-form solutions. As a result, standard convex-optimization methods, such as interior point, could be used in order to solve the problem. The interior point method is known to have a polynomial time complexity [40].

3.3 Simulations and Numerical Results

The results illustrated in the following figures are extracted by means of Monte Carlo simulations. Specifically, in each test, different channel gains are generated and results are obtained from the average performance of the tests. The purpose of the simulations is to investigate the effect of different parameters on the latency threshold of URLLC users. We evaluate the performance of the proposed optimization problem by comparing its results with selected benchmarks. In particular, its solution is compared with those when the whole bandwidth is equally allocated among users and when the whole cloud capacity is equally allocated among users respectively.

We consider that URLLC and mMTC users are uniformly distributed in a circular area with radius R . The random gain h_k under Rayleigh fading follows exponential distribution with mean being 2. Initially, we consider a system which consists of $K = 10$ users in total, where $U = 5$ of them belong to the URLLC subset and the rest $M = 5$ to the mMTC one. In table 3.1, the significant parameters of the system are presented.

Parameter	Value	Parameter	Value
B	1MHz	L_k	50 kbits
F	10GHz	C_k	100 CPU cycles
N_o	-174dBm/Hz	$p_{\max,k}$	0.5 W
a	2	R	500m

Table 3.1: Simulation parameters

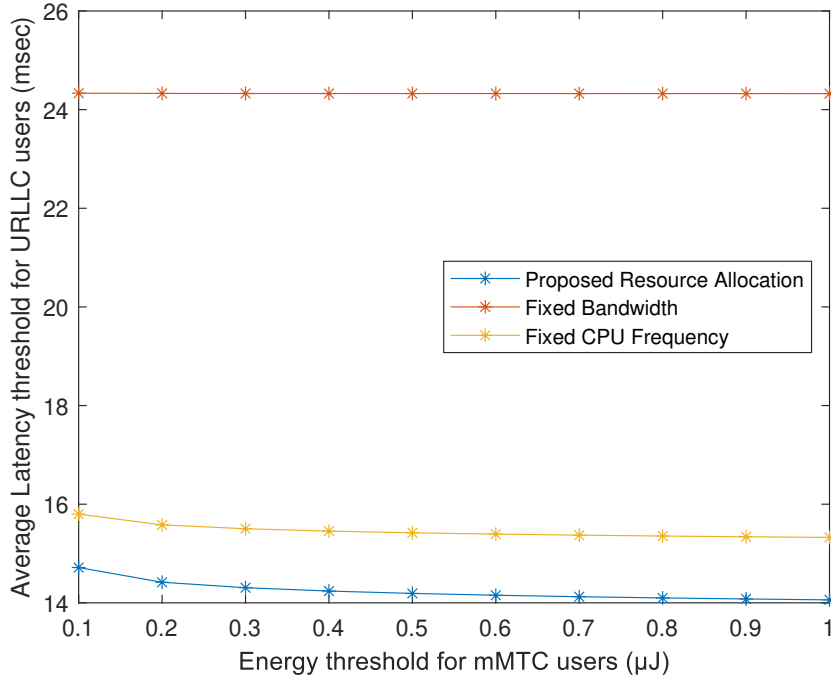


Figure 3.1: Average optimal URLLC latency threshold as a function of mMTC energy consumption threshold for different channel gain coefficients ($E_u = 0.1\text{mJ}$, $T_m = 0.1\text{sec}$)

In Figures 3.1 and 3.2 we represent the average achieved latency threshold of URLLC users, calculated for different channel coefficients and over different values of the energy consumption threshold of mMTC users and of URLLC users respectively. In Figure 3.1 the proposed optimization significantly outperforms the fixed bandwidth scheme by 71% and the fixed CPU frequency scheme by 9%, while in Figure 3.2 by 70% and 15% respectively. The variable CPU frequency seems to have a small effect on reducing URLLC latency. This happens because the computation capacity of the MEC server is huge, resulting in very low execution time. For the given increasing energy threshold values, the latency threshold decreases. In order to achieve these values, the optimization algorithm for the proposed optimization problem deploys the whole bandwidth and CPU capacity. But more specifically, it allocates more CPU capacity and bandwidth to the URLLC users in order to achieve the smallest possible latency threshold, satisfying at the same time the constraints for the mMTC users. Furthermore, both sets of devices transmit with power lower than their maximum.

In Figure 3.3 we investigate the average optimal URLLC latency threshold as a function of the mMTC latency threshold, for different channel gain coefficients. Once again, the proposed resource allocation is superior to those with fixed bandwidth and CPU clock speed. The fixed CPU frequency scheme has an 16% increase in delay and the fixed bandwidth scheme has an 86% increase in delay in comparison with the proposed scheme. It is worth noticing that, with the proposed resource allocation, when allowing LE devices to be more delay-tolerant, we can achieve a latency threshold difference between mMTC and URLLC devices of two orders of magnitude. Once more, the whole bandwidth and CPU capacity are deployed for the proposed resource allocation.

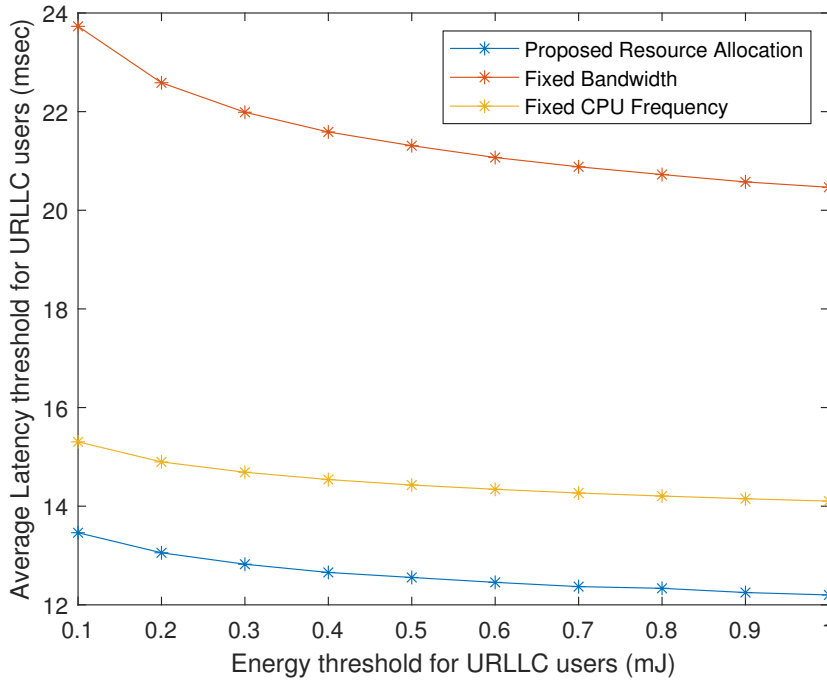


Figure 3.2: Average optimal URLLC latency threshold as a function of URLLC energy consumption threshold for different channel gain coefficients ($E_m = 1\mu\text{J}$, $T_m = 0.1\text{sec}$)

Figure 3.4 depicts the average optimal URLLC latency threshold as a function of URLLC (or mMTC) devices for different channel gain coefficients. As the number of mobile user devices increases, the delay increases, as expected, owing to the additional complexity. Similarly to the results of previous simulations, the proposed optimization scheme outmatches the other two. The observed increase in delay from the proposed optimization is 3% for fixed CPU capacity and 44% for fixed bandwidth.

Finally, Figure 3.5 shows the optimal average URLLC latency threshold as a function of task size of devices for different channel gain coefficients. Here we observe that for increasing values of task size L , leads to higher attainable URLLC latency threshold. This is reasonable since the offloading delay increases proportionally to the size of input data. Moreover, in our proposed and fixed CPU capacity resource allocation schemes, delay increases much smoother as the size of users' tasks increases, compared to the fixed bandwidth case. A reasonable explanation to this is the ability of the interior-point algorithm to adjust the allocated bandwidth to URLLC users in order to offset the influence of the increasing task size to the offloading delay. The proposed topology scales better than the fixed allocated MEC capacity and fixed bandwidth scheme, resulting in a 15% and a 89% decrease respectively.

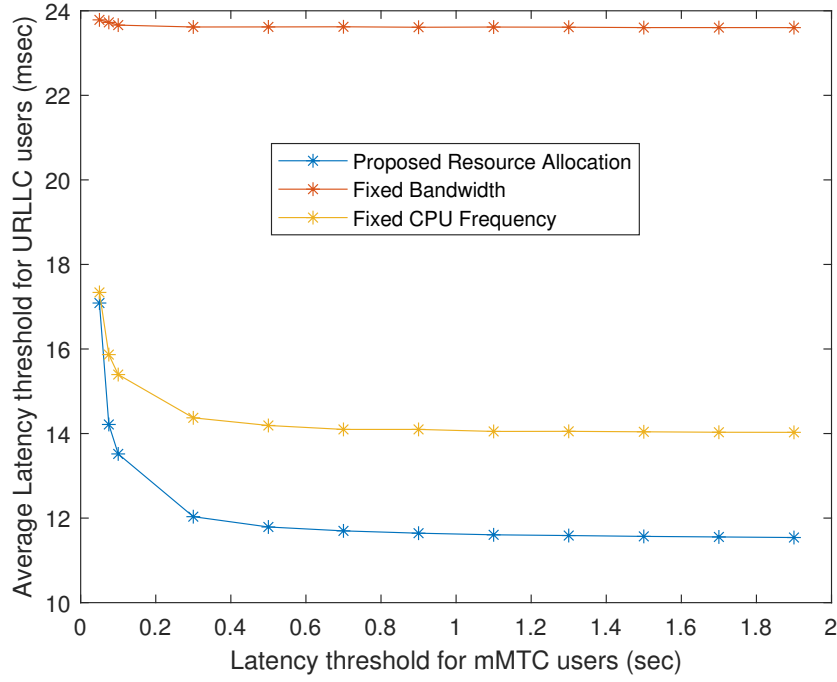


Figure 3.3: Average optimal URLLC latency threshold as a function of mMTC latency threshold for different channel gain coefficients ($E_u = 0.1\text{mJ}$, $E_m = 1\mu\text{J}$)

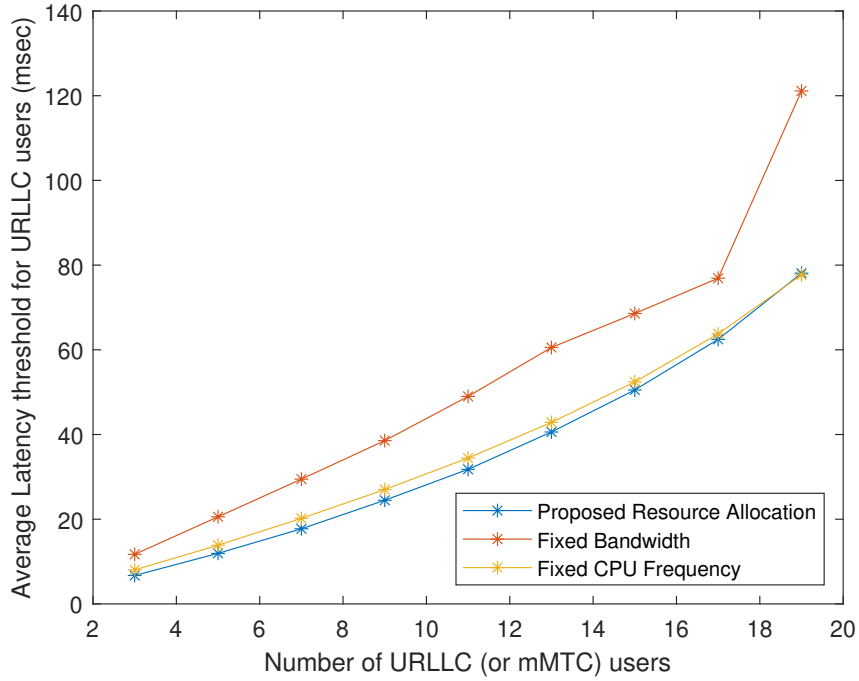


Figure 3.4: Average optimal URLLC latency threshold as a function of URLLC (or mMTC) devices for different channel gain coefficients ($E_u = 0.1\text{mJ}$, $E_m = 1\mu\text{J}$, $T_m = 0.1\text{sec}$)

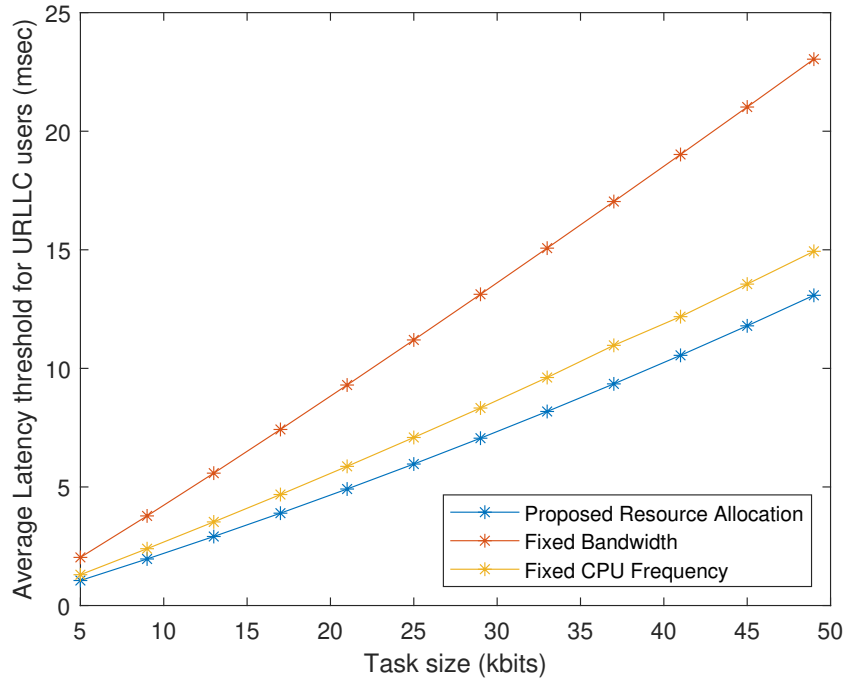


Figure 3.5: Average optimal URLLC latency threshold as a function of task size of devices for different channel gain coefficients ($E_u = 0.1\text{mJ}$, $E_m = 1\mu\text{J}$, $T_m = 0.1\text{sec}$)

Chapter 4

mMTC Sum Energy Consumption Minimization with NOMA

4.1 Introduction

Wireless networks, before the emergence of the 5G New Radio, allocate radio resources to users based on the orthogonal multiple access (OMA) principle. Time division multiple access (TDMA), frequency division multiple access (FDMA), and code division multiple access (CDMA) were some of the most commonly used orthogonal multiple access techniques in previous generations of cellular networks. Regardless of the fact that OMA approaches can achieve high system performance even with simple receivers, due to no mutual interference among users in an ideal situation, they are unable to solve the rising issues posed by increasing demands in 5G networks and beyond. The concept of nonorthogonal multiple access (NOMA) appears as a solution to enhance spectral efficiency while allowing some degree of multiple access interference at receivers. According to 3GPP, by using NOMA scheme for mMTC and URLLC applications, the number of user connections can be increased by 5 and 9 times, respectively [41], [42].

4.2 System Model

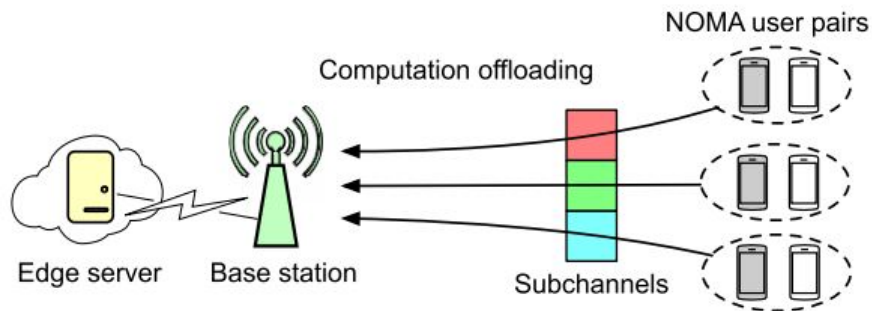


Figure 4.1: NOMA user pairing for MEC

A single cell MEC offloading scenario is considered with K users, which are served by a single BS equipped with a MEC server. The set of users consists of both URLLC and mMTC users. In terms of simplicity, we consider that the number of URLLC devices is equal to the number of mMTC devices, i.e. the cell consists of $2N$ users, where $N = K/2$. Let $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$ be the set of all N URLLC users and $\mathcal{M} = \{m_1, m_2, \dots, m_N\}$ be the set of all N mMTC users in the network.

The goal is to pair the users into N NOMA user pairs, where each pair consists of one mMTC user m and one URLLC user u . Each NOMA pair operates on a frequency resource block (RB) which is orthogonal to other frequency RBs. In order to ensure reliable transmission for URLLC devices, for the pairing process, we consider that URLLC users have already been selected by the scheduler to transmit in RBs. An example of three NOMA pairs is presented in Figure 4.1. In more detail, the total system bandwidth is equally divided into N frequency RBs, denoted as $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$, each with a bandwidth of $B_s = B/N$ and each assigned to only one URLLC user and one mMTC user.

The users perform all computation at the MEC server. Due to the MEC server's huge computation capacity, which results in very short data processing time, and the small sizes of computation results, the time for execution of computations at the cloud and downloading from the BS is insignificant compared to the time for mobile offloading. Therefore, the total offloading delay is approximated to the time for data uploading to the edge cloud.

We define as $p_{u,s}$ and $p_{m,s}$ the transmission power of a URLLC user and an mMTC user respectively on a resource block s , $\forall s \in \mathcal{S}$, and d_u , d_m their distance from the base station. An illustration of the designed system can be seen in Figure 4.2. Our priority is to ensure low energy consumption of mMTC devices and to serve reliably and interference-free the URLLC users within a strict delay constraint. Therefore, we assume that device m is the strong one and it has to be decoded first according to the scheme of NOMA. Specifically, in every pair, the mMTC signal is decoded first, thus receiving interference from the URLLC signal, and then it is subtracted from the received signal, following the Successive Interference Cancellation (SIC) scheme. After that, the URLLC signal is decoded without any interference.

We define the channel gains as $g_{u,s} \triangleq |h_{u,s}|^2 d_u^{-a}$ and $g_{m,s} \triangleq |h_{m,s}|^2 d_m^{-a}$ respectively, where $h_{u,s}$ and $h_{m,s}$ are the exponential channel gains (corresponding to Rayleigh fading) of the u -th and m -th user on sub-channel s , and a is the path loss exponent. Each user u and each user m is required to offload L_u and L_m bits of input data respectively. We assume that the transmit power of each device is variable, i.e., $0 < p_{u,s} \leq p_{\max,u}$ and $0 < p_{m,s} \leq p_{\max,m}$. Let $a_{m,s}$ denote the subchannel allocation decision of user m , i.e.,

$$a_{m,s} = \begin{cases} 1 & \text{if user } m \text{ is assigned to subchannel } s, \\ 0 & \text{otherwise} \end{cases} \quad (4.1)$$

such that, $\sum_{s \in \mathcal{S}} a_{m,s} = 1$, $\forall m \in \mathcal{M}$ and $\sum_{m \in \mathcal{M}} a_{m,s} = 1$, $\forall s \in \mathcal{S}$. Since the bandwidth of each subchannel is considered constant, we define $\tilde{g}_{u,s} \triangleq \frac{g_{u,s}}{B_s N_o}$ and $\tilde{g}_{m,s} \triangleq \frac{g_{m,s}}{B_s N_o}$ as the normalized channel gains, with N_o being the power spectral density of the additive white gaussian noise (AWGN). Then, for a device u assigned to subchannel s , the uplink transmission data rate follows

$$r_{u,s} = B_s \log_2 (1 + p_{u,s} \tilde{g}_{u,s}), \quad \forall u \in \mathcal{U}, \quad \forall s \in \mathcal{S} \quad (4.2)$$

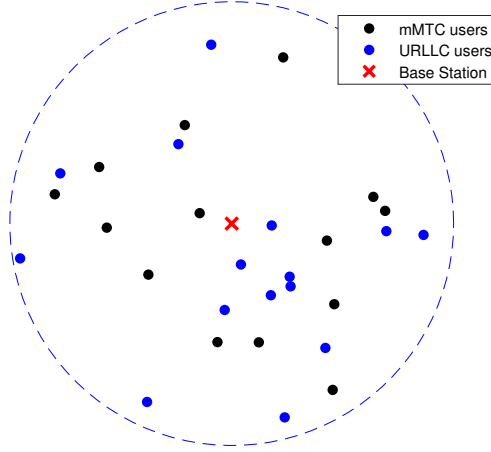


Figure 4.2: System design of distance setting

and if user m accesses it, it follows

$$r_{m,s} = B_s \log_2 \left(1 + \frac{p_{m,s} \tilde{g}_{m,s}}{p_{u,s} \tilde{g}_{u,s} + 1} \right), \quad \forall m \in \mathcal{M}, \quad \forall s \in \mathcal{S} \quad (4.3)$$

The time required to offload the data to the cloud is

$$t_{\text{off},u} = \frac{L_u}{r_{u,s}}, \quad \forall u \in \mathcal{U}, \quad \forall s \in \mathcal{S} \quad (4.4)$$

$$t_{\text{off},m} = \frac{L_m}{r_{m,s}}, \quad \forall m \in \mathcal{M}, \quad \forall s \in \mathcal{S} \quad (4.5)$$

where

$$r_m = \sum_{s \in \mathcal{S}} a_{m,s} r_{m,s}, \quad \forall m \in \mathcal{M} \quad (4.6)$$

As in the previous chapter, the latency requirements of URLLC and mMTC users should satisfy $t_{\text{off},u} \leq T_u, \forall u \in \mathcal{U}$ and $t_{\text{off},m} \leq T_m, \forall m \in \mathcal{M}$ respectively, where $T_u \ll T_m$. Finally, the energies consumed for the offload can be expressed as follows:

$$E_{\text{off},u} = t_{\text{off},u} p_{u,s}, \quad \forall u \in \mathcal{U} \quad (4.7)$$

$$E_{\text{off},m} = t_{\text{off},m} \sum_{s \in \mathcal{S}} a_{m,s} p_{m,s}, \quad \forall m \in \mathcal{M} \quad (4.8)$$

and they have to satisfy $E_{\text{off},u} \leq E_u, \forall u \in \mathcal{U}$ and $E_{\text{off},m} \leq E_m, \forall m \in \mathcal{M}$ respectively, where $E_m \ll E_u$.

4.3 Problem Formulation

With the objective of prolonging the lifetime of mMTC devices, we consider the problem of their total transmission energy minimization, under offloading delay, energy, transmission power restrictions for both set of users and mMTC subchannel assignment constraints. Since the objective is to find the optimal minimum mMTC energy consumption, there is no need in constraining

it, meaning that constraint $E_{\text{off},m} \leq E_m$ can be omitted. However, for each user u , the energy consumption should be less than a threshold E_u , as mentioned in the previous section. Also, the transmission time of devices is restricted by a latency threshold T_n , which is stricter for URLLC devices. The problem jointly considers the subchannel assignment and power control. It is formulated as follows

$$\begin{aligned}
& \min_{\{\mathbf{p}_u, \mathbf{p}_m, \mathbf{A}\}} \sum_{m \in \mathcal{M}} E_{\text{off},m} \\
& \text{s.t.} \\
& C1: t_{\text{off},m} \leq T_m \Rightarrow \frac{L_m}{r_m} \leq T_m, \quad \forall m \in \mathcal{M} \\
& C2: t_{\text{off},u} \leq T_u \Rightarrow \frac{L_u}{r_u} \leq T_u, \quad \forall u \in \mathcal{U} \\
& C3: E_{\text{off},u} \leq E_u \Rightarrow \frac{L_u}{r_u} p_{u,s} \leq E_u, \quad \forall u \in \mathcal{U} \\
& C4: p_{m,s} \leq p_{\max,m}, \quad p_{m,s} > 0, \quad \forall m \in \mathcal{M}, \quad \forall s \in \mathcal{S} \\
& C5: p_{u,s} \leq p_{\max,u}, \quad p_{u,s} > 0, \quad \forall u \in \mathcal{U}, \quad \forall s \in \mathcal{S} \\
& C6: \sum_{s \in \mathcal{S}} a_{m,s} = 1, \quad \forall m \in \mathcal{M} \\
& C7: \sum_{m \in \mathcal{M}} a_{m,s} = 1, \quad \forall s \in \mathcal{S} \\
& C8: a_{m,s} \in \{0, 1\}, \quad \forall m \in \mathcal{M}, \quad \forall s \in \mathcal{S}
\end{aligned} \tag{P2}$$

where $E_{\text{off},m}$ in the objective function is given by (4.8), $\mathbf{A} = [\mathbf{a}_{m,1}, \mathbf{a}_{m,2}, \dots, \mathbf{a}_{m,s_N}]$ is a matrix containing mMTC user allocation vectors, while matrices $\mathbf{p}_u = [\mathbf{p}_{u,1}, \mathbf{p}_{u,2}, \dots, \mathbf{p}_{u,s_N}]$ and $\mathbf{p}_m = [\mathbf{p}_{m,1}, \mathbf{p}_{m,2}, \dots, \mathbf{p}_{m,s_N}]$ contain URLLC and mMTC transmission power vectors respectively. Inequality constraint $C1$ reflects the mMTC offloading delay restriction. $C2$ and $C3$ limit the delay and energy of URLLC users. Constraints $C4$ and $C5$ restrict the transmission power of each user. Constraint $C6$ ensures that each mMTC user can have access to only one subchannel, while constraint $C7$ ensures that each subchannel is paired to only one user. Finally, constraint $C8$ restricts $a_{m,s}$ to binary choice.

4.4 Proposed Solution

Problem P2 is a non-convex combinatorial problem, which is hard to solve. To tackle the complexity issue, a decomposition strategy is followed. In particular, the initial problem is decomposed into two sub-problems so that each of them can be solved using a suitable approach. The decomposition leads to a two-phase iterative algorithm, where each sub-problem is initialised with the optimal solution of the other sub-problem. The first sub-problem addresses the subchannel assignment of mMTC devices, while the second concerns the power allocation problem in every subchannel. It should be noted that, since the allocation of the MEC computing resources is neglected, the power allocation problem is independent for each pair. For the power allocation sub-problem, the closed form solutions are obtained. The two subproblems are iteratively solved until convergence.

4.4.1 Subchannel Assignment for a Fixed Power Allocation

For the pairing process, as mentioned above, URLLC users have been already selected by the scheduler to transmit in RBs. mMTCs are in need of resources but they are not sorted in any particular order. Assuming that the devices transmit with given constant powers $p_{u,s}$ and $p_{m,s}$, the assignment problem of mMTC users to RBs/URLLC users is mathematically expressed as:

$$\begin{aligned}
& \min_{\mathbf{A}} \quad \sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{S}} a_{m,s} C_{m,s} \\
& \text{s.t.} \\
& C1: \sum_{m \in \mathcal{M}} a_{m,s} = 1, \quad \forall s \in \mathcal{S} \\
& C2: \sum_{s \in \mathcal{S}} a_{m,s} = 1, \quad \forall m \in \mathcal{M} \quad (\mathbf{P3}) \\
& C3: a_{m,s} \in \{0, 1\}, \quad \forall m \in \mathcal{M}, \quad \forall s \in \mathcal{S}
\end{aligned}$$

where $\mathbf{A} = [\mathbf{a}_{m,1}, \mathbf{a}_{m,2}, \dots, \mathbf{a}_{m,s_N}]$ is a $N \times N$ matrix containing mMTC user allocation vectors and $C_{m,s}$ is the cost of assigning the m -th mMTC user to the s -th subchannel. Specifically, $C_{m,s}$ reflects the energy consumption of an mMTC user m for the offloading in a subchannel s and interfered by a URLLC user u and we define the square $N \times N$ cost matrix C such that the element of the m -th row and the s -th column, i.e, $C_{m,s}$ is set to

$$C_{m,s} = \frac{p_{m,s} L_m}{B_s \log_2 \left(1 + \frac{p_{m,s} \tilde{g}_{m,s}}{p_{u,s} \tilde{g}_{u,s} + 1} \right)} \quad (4.9)$$

The optimization problem P3 is a typical assignment problem. We could obtain an optimal user-pairing combination by using an exhaustive search algorithm. For the exhaustive pairing, all possible pairings are examined in order to determine the combination of pairs with the minimum cost. However, the complexity of the exhaustive search algorithm is $O(N!)$, where N is the number of users to be assigned to N subchannels, and thus it is computationally expensive for a large number of users.

Therefore, we utilize the Hungarian Algorithm for finding the corresponding optimal solution of the above problem. The Hungarian algorithm is an accurate combinatorial optimization method used for solving a two-sided one-to-one matching problem and well known for its use in solving resource allocation problems [43]. It has a complexity of $O(N^3)$ which is much better than for the exhaustive search algorithm. It finds the optimal solution by minimizing a cost associated with a set of pairs. The Hungarian method is presented in steps in Algorithm 1. Using Algorithm 1 we can easily get the optimal solution to the assignment problem. In order to clearly demonstrate how the Hungarian algorithm works, an example is presented in Appendix A.

4.4.2 Power Allocation for a Fixed Subchannel Assignment

After obtaining the assignment/pairing of mMTC users to subchannels/URLLC users, assuming transmission with fixed power, we should now obtain the optimal power allocation solution of each NOMA pair. As mentioned above, the power allocation problem is independent for each

Algorithm 1: Hungarian algorithm for sub-channel assignment

-
- Input:** Cost Matrix C
Output: The optimal sub-channel assignment matrix H^*
- 1 Row Reduction: find the minimum value of each row and subtract it from every value of the that row
 - 2 Column Reduction: find the minimum value of each column and subtract it from every value of the that column
 - 3 Cover all zeros in the rows and columns using a minimum number of horizontal or vertical lines
 - 4 Optimality check: (a) If the minimum number of covering lines is N , an optimal assignment of zeros which don't lie in the same row or column is possible. The algorithm stops. (b) If the minimum number of covering lines is less than N , then continue with Step 5.
 - 5 Find the smallest element that is not covered by a line. Subtract this element from each uncovered row, and then add it to each covered column. Return to Step 2.
-

pair of mobile devices. Let consider a subchannel s of interest and u and m the URLLC and mMTC users occupying subchannel s respectively. Recall that the objective is to minimize the energy consumption of mMTC users for data offloading. Then, the power allocation optimization problem for a subchannel s can be written as

$$\begin{aligned}
& \min_{\{p_u, p_m\}} \quad \frac{p_m L_m}{B_s \log_2 \left(1 + \frac{p_m \tilde{g}_m}{p_u \tilde{g}_u + 1} \right)} \\
& \text{s.t.} \\
& C1 : t_{\text{off},m} \leq T_m \Rightarrow \frac{L_m}{B_s \log_2 \left(1 + \frac{p_m \tilde{g}_m}{p_u \tilde{g}_u + 1} \right)} \leq T_m \\
& C2 : t_{\text{off},u} \leq T_u \Rightarrow \frac{L_u}{B_s \log_2 (1 + p_u \tilde{g}_u)} \leq T_u \quad (\mathbf{P4}) \\
& C3 : E_{\text{off},u} \leq E_u \Rightarrow \frac{p_u L_u}{B_s \log_2 (1 + p_u \tilde{g}_u)} \leq E_u \\
& C4 : p_u \leq p_{\max,u}, \quad p_u > 0 \\
& C5 : p_m \leq p_{\max,m}, \quad p_m > 0
\end{aligned}$$

Problem P4 is non-convex, due to the coupling of the transmission powers in the objective function. However, after some transformations, it is possible to obtain the closed-form power controls for users's uplink data transmissions. Initially, by observing the impact of the optimization variables on the objective function, we provide the following two propositions.

Proposition 1. *The objective function is monotonically increasing with respect to p_u .*

Proof. We define a function f :

$$f(x) = \frac{l_1}{\log_2 \left(1 + \frac{l_2}{l_3 x + 1} \right)}, \quad (4.10)$$

where $0 < x = p_u \leq p_{\max,u}$, $l_1 = \frac{p_m L_m}{B_s} > 0$, $l_2 = p_m \tilde{g}_m > 0$ and $l_3 = \tilde{g}_u > 0$. We calculate its derivative :

$$f'(x) = \frac{l_1 l_2 l_3 \ln(2)}{(1 + l_3 x)^2 \left(1 + \frac{l_2}{1 + l_3 x} \right) \left[\ln \left(1 + \frac{l_2}{1 + l_3 x} \right) \right]^2} > 0, \quad (4.11)$$

Since all terms are positive, the derivative is positive for every $x \in (0, p_{\max, u}]$. Therefore, $f(p_u)$ is an increasing function w.r.t. p_u and the proof is completed. \square

Proposition 2. *The objective function is monotonically increasing with respect to p_m .*

Proof. We define a function

$$f(x) = \frac{w_1 x}{\log_2(1 + w_2 x)}, \quad (4.12)$$

where $0 < x = p_m \leq p_{\max, m}$, $w_1 = \frac{L_m}{B_s} > 0$ and $w_2 = \frac{\tilde{g}_m}{p_u \tilde{g}_u + 1} > 0$. Then, the first order derivative of $f(x)$ is given by

$$f'(x) = \frac{g_1(x) - g_2(x)}{[\log_2(1 + w_2 x)]^2}, \quad (4.13)$$

while $g_1(x)$ and $g_2(x)$ are given by

$$g_1(x) = w_1 \log_2(1 + w_2 x) \quad (4.14)$$

and

$$g_2(x) = \frac{w_1 w_2 x}{(\ln 2)(1 + w_2 x)}, \quad (4.15)$$

respectively. Then, the first order derivatives of $g_1(x)$ and $g_2(x)$ are

$$g_1'(x) = \frac{w_1 w_2}{(\ln 2)(1 + w_2 x)} \quad (4.16)$$

and

$$g_2'(x) = \frac{w_1 w_2}{(\ln 2)(1 + w_2 x)^2}, \quad (4.17)$$

respectively. Note that $g_1'(x) > g_2'(x)$, since $\frac{g_1'(x)}{g_2'(x)} = 1 + w_2 x > 1$. Moreover, we observe that $g_1(0) = g_2(0)$, thus $g_1(x) > g_2(x)$ for $x \in (0, p_{\max, m}]$ and consequently, $f'(x) > 0$ for $x \in (0, p_{\max, m}]$, meaning that $f(x)$ is an increasing function w.r.t. p_m . \square

From Propositions 1 and 2 we conclude that, since the objective function is increasing w.r.t. the transmission powers, the optimal mMTC energy consumption is given by the minimum transmission powers p_u and p_m that satisfy the constraints. The optimal p_u can be obtained as in the following lemma.

Lemma 2. *The optimal p_u of problem P_4 is given by*

$$p_u^* = \frac{1}{\tilde{g}_u} (2^{\frac{L_u}{B_s T_u}} - 1) \quad (4.18)$$

when constraint C_4 is satisfied and when

$p_u^* < \frac{-b \ln(2) - a \mathcal{W}_{-1}(c)}{a b \ln(2)}$, where $a = \tilde{g}_u$, $b = \frac{L_u}{E_u B_s}$, $c = -\frac{2^{\frac{b}{a}} b \ln(2)}{a}$ and $\mathcal{W}_{-1}(\cdot)$ denotes the secondary branch of the Lambert W function.

Proof. We observe that by manipulating constraint C_2 , we arrive at

$$p_u \geq z_1, \quad (4.19)$$

where $z_1 = \frac{1}{\tilde{g}_u} (2^{\frac{L_u}{B_s T_u}} - 1)$, which yields a lower bound of p_u . Furthermore, constraint C_3 can be written in the following form

$$p_u L_u - E_u B_s \log_2(1 + p_u \tilde{g}_u) \leq 0 \quad (4.20)$$

It is easy to verify that the above function is convex w.r.t p_u , thus it has at most two roots r_1, r_2 . The roots are

$$r_1 = 0, \quad (4.21)$$

$$r_2 = \frac{-b \ln(2) - a \mathcal{W}_{-1}(c)}{a b \ln(2)}, \quad (4.22)$$

where $a = \tilde{g}_u$, $b = \frac{L_u}{E_u B_s}$, $c = -\frac{2^{\frac{b}{a}} b \ln(2)}{a}$ and \mathcal{W}_{-1} is secondary branch of the Lambert W function. In order (4.20) to be satisfied, $p_u \in [r_1, r_2]$ should hold. Thus, constraints $C2$, $C3$ and $C4$ provide the upper and lower bounds of p_u . Following that, it is straightforward to show that conditions $z_1 < r_2$ and $z_1 < p_{\max, u}$ should be satisfied in order problem P4 to be feasible. When satisfied, it follows that $p_u \geq z_1$. From Proposition 1, we derive that in order to minimize $E_{\text{off}, m}$, p_u should be the minimum. Therefore, the optimal p_u^* is given by

$$p_u^* = z_1 = \frac{1}{\tilde{g}_u} (2^{\frac{L_u}{B_s T_u}} - 1) \quad (4.23)$$

and the proof is completed. \square

After obtaining the optimal solution for p_u , next, the optimal solution for p_m is given by Lemma 3, when the aforementioned conditions in Lemma 2 and constraint $C5$ hold for feasibility.

Lemma 3. *The optimal p_m of problem P4 is given by*

$$p_m^* = \frac{1}{\tilde{g}_m} (2^{\frac{L_m}{B_s T_m}} - 1) 2^{\frac{L_u}{B_s T_u}} \quad (4.24)$$

Proof. From Proposition 2, we conclude that in order to minimize $E_{\text{off}, m}$, p_m should be the minimum. Following that, we can obtain the minimum transmission power requirement p_m when the maximum mMTC latency constraint is satisfied, meaning that constraint $C1$ holds with equality. This can be easily verified, since $t_{\text{off}, m}$ is a decreasing function w.r.t p_m , i.e.

$$\frac{dt_{\text{off}, m}(p_m)}{dp_m} = -\frac{L_m \tilde{g}_m \ln(2)}{B_s (1 + p_u \tilde{g}_u) \left(1 + \frac{p_m \tilde{g}_m}{1 + p_u \tilde{g}_u}\right) \left[\ln \left(1 + \frac{p_m \tilde{g}_m}{1 + p_u \tilde{g}_u}\right)\right]^2} < 0. \quad (4.25)$$

Then, the optimal p_m^* is given by

$$p_m^* = \frac{1}{\tilde{g}_m} (2^{\frac{L_m}{B_s T_m}} - 1) (p_u^* \tilde{g}_u + 1) = \frac{1}{\tilde{g}_m} (2^{\frac{L_m}{B_s T_m}} - 1) 2^{\frac{L_u}{B_s T_u}} \quad (4.26)$$

and the proof is completed. \square

Finally, the minimum energy consumption of an mMTC user m , offloading in a subchannel s and interfered by a URLLC user u is given by

$$E_{\text{off}, m}^* = \frac{p_m^* L_m}{B_s \log_2 \left(1 + \frac{p_m^* \tilde{g}_m}{p_u^* \tilde{g}_u + 1}\right)} = p_m^* T_m \quad (4.27)$$

4.4.3 Iterative Algorithm Description

After analysing the approaches followed to solve the sub-problems of channel assignment and power allocation, now in order to solve problem P2, we use an iterative approach in which power allocation and subchannel assignment are performed iteratively to achieve the optimal solution. The process is summarized in Algorithm 2, where i is the iteration index. $\mathbf{P}(u)$, $\mathbf{P}(m)$ and \mathbf{A} are vectors related with the transmission powers of URLLCs, mMTCs and subchannel assignment of mMTCs respectively.

First, initial values for $\mathbf{P}(u)$, $\mathbf{P}(m)$ and \mathbf{A} are defined. In particular, in the beginning we set the transmission power of each device in every subchannel equal to its maximum. Next, we find the optimal matching \mathbf{A}^i with previous power vectors $\mathbf{P}^{i-1}(u)$ and $\mathbf{P}^{i-1}(m)$. Then, for obtained subchannel vectors in iteration i , if the aforementioned feasibility conditions analysed in the previous section hold, we compute the optimal power vectors. Otherwise, the algorithm stops, since there is no feasible solution to Problem P4. This process is executed iteratively until convergence. Specifically, the iteration is stopped when the matching vectors do not change, i.e. \mathbf{A}^{i-1} is identical to \mathbf{A}^i , and when $\|\mathbf{P}^i(m) - \mathbf{P}^{i-1}(m)\|$ and $\|\mathbf{P}^i(u) - \mathbf{P}^{i-1}(u)\|$ are smaller than e , where $0 < e \ll 1$.

Algorithm 2: Joint subchannel and power allocation iterative algorithm

```

1: Initialize:  $\mathbf{P}^0(u)$ ,  $\mathbf{P}^0(m)$ ,  $\mathbf{A}^0$  and set  $i = 1$ 
2: Repeat:
3:   Step 1:
4:     For fixed  $\mathbf{P}^{i-1}(u)$  and  $\mathbf{P}^{i-1}(m)$ , calculate cost matrix  $\mathbf{C}$  from (4.9)
5:     Compute optimal assignment  $\mathbf{A}^i$  using Algorithm 1
6:   Step 2:
7:   for each subchannel  $s = 1, \dots, N$ 
8:     Calculate:  $r_2$  and  $z_1$  using (4.22) and (4.23)
9:     if  $z_1 < r_2$  and  $z_1 < p_{\max,u}$ 
10:       $p_{u,s}^* = z_1$ 
11:      obtain  $p_{m,s}^*$  from (4.24)
12:      if  $p_{m,s}^* > p_{\max,m}$ 
13:         $p_{u,s}^* = \text{NaN}$ ,  $p_{m,s}^* = \text{NaN}$ 
14:        break repeat
15:      end if
16:    else
17:       $p_{u,s}^* = \text{NaN}$ ,  $p_{m,s}^* = \text{NaN}$ 
18:      break repeat
19:    end if
20:  end for
21:   $\mathbf{P}^i(u) \leftarrow [p_{u,1}^*, p_{u,2}^*, \dots, p_{u,N}^*]$ 
22:   $\mathbf{P}^i(m) \leftarrow [p_{m,1}^*, p_{m,2}^*, \dots, p_{m,N}^*]$ 
23:   $i = i + 1$ 
24: Until:  $\|\mathbf{P}^i(m) - \mathbf{P}^{i-1}(m)\| \leq e$  AND  $\|\mathbf{P}^i(u) - \mathbf{P}^{i-1}(u)\| \leq e$  AND  $\mathbf{A}^i \equiv \mathbf{A}^{i-1}$ 

```

4.5 Simulations and Numerical Results

The same performance evaluation as in Chapter 3 is considered, which means that most of the following figures are extracted by means of Monte Carlo simulations. In each test, the distance of the devices from the BS and the gain h are different, resulting in different channel gains, thus different matchings. The results are derived from the average performance of the tests. The purpose of the simulations is to investigate the effect of various parameters on the sum of all mMTCs' energy consumption. We will compare the proposed solution with the solution resulting from the exhaustive search and the random selection for the assignment of mMTC users to subchannels.

MMTC and URLLC devices are uniformly distributed in a circular area with radius R , as illustrated in Figure 4.2. Under Rayleigh fading, the random gain h follows an exponential distribution with mean being 2 and the pathloss exponent is equal to 2. We assume that all users send the same amount of input data L . In table 4.1, some other the significant parameters of the system are presented.

Figure 4.3 illustrates the convergence behavior, for different number of users, of Algorithm 2 that is used to find the optimal subchannel and power allocation by iteratively solving problems P3 and P4. It should be mentioned that Monte Carlo simulations are not used in this case, therefore the system's parameters are evaluated using established channel gains. Also, the resulted energy in every iteration is normalized by the resulted maximum energy from all iterations. As seen in the figure, the proposed algorithm converges very fast, even for an increasing number of users. In particular, in this case, the algorithm finds the optimal solution in the third iteration and stops in the fourth one, since it has converged.

Note that all the following figures are base-10 logarithmic scaled on the y-axis. In Figure 4.4 we have assumed that the number of users in all subchannels is 20, 10 URLLCs and 10 mMTCs, and we compare the resulted average and minimum achieved transmission power of the mMTC and URLLC users in every subchannel, for different channel coefficients, using the Hungarian method. It is obvious that, for the given parameters, there is a big difference between the transmission powers of devices. In detail, URLLC mobile device users transmit with much higher power in every subchannel. This is reasonable, since URLLC users are not optimally assigned to subchannels so that the energy consumption of mMTC devices, and thus the transmission power of URLLCs, to be the minimum.

Parameter	Value	Parameter	Value
R	500m	N_o	-174dBm/Hz
B	1MHz	L	50 kbits
$p_{\max,u}$	24 dBm	$p_{\max,m}$	21 dBm

Table 4.1: Simulation parameters

In Figure 4.5 we present the average minimum sum energy consumption of the total mMTC

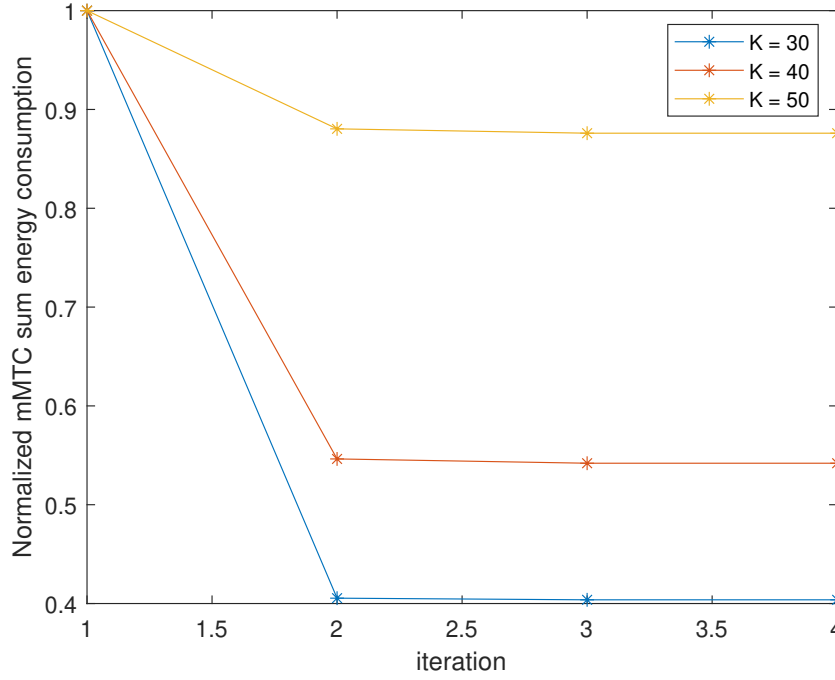


Figure 4.3: Convergence behavior of Algorithm 2 for different numbers of total users ($E_u = 0.1\text{J}$, $T_u = 0.05\text{sec}$, $T_m = 1.5\text{sec}$)

users in the system, for different channel coefficients, as a function of the increasing number of resource blocks and consequently of devices. We compare the proposed solution using the Hungarian method and the exhaustive search method. Note that we do not provide results for greater number of devices/RBs due to the high complexity of the exhaustive search algorithm. As it was expected, when traffic increases, the achievable total energy of mMTC users increases too. It is also worth noticing that the Hungarian Algorithm has the same performance with that of the exhaustive search approach, meaning that it provides optimal results in less time.

In order to reduce the complexity and the running time of the simulations, the rest of the figures are extracted for a total number of 10 users, where $U = 5$ of them belong to the URLLC subset and $M = 5$ to the mMTC one. Figure 4.6 shows the average minimum sum energy consumption of mMTC users as a function of their latency threshold and for different channel gain coefficients. We can observe that there is a trade-off between offloading energy consumption and offloading delay, i.e. the mMTC energy decreases as their latency threshold increases. This can be easily verified from the analysis in the previous section, since the mMTC energy consumption, as proven, is an increasing function w.r.t. mMTC transmission power and the mMTC offloading delay, on the other hand, is a decreasing function w.r.t. mMTC transmission power. So, the increase in mMTC latency threshold causes a decrease in mMTC power consumption and subsequently in mMTC energy consumption. Furthermore, from (4.18) it is evident that the URLLC transmission power is not affected by the mMTC latency threshold, and consequently the different values of T_m does not have an impact on the URLLC power transmission.

Finally, in figures 4.7 and 4.8 we present the average minimum sum energy consumption of mMTC users as a function of the latency threshold of URLLC users and of the size of the task the

devices have to offload, respectively. In these figures, the effectiveness of the proposed approach is also compared to the performance of a random selection approach, in which mMTC users are served by randomly selecting resource blocks, provided that their requirements are satisfied. It is obvious that both the average sum energy consumption resulting from the Hungarian approach and the exhaustive search method are superior to that of the random assignment approach. Once again, it appears to be a trade-off between offloading energy consumption and offloading delay. In more detail, when URLLC latency threshold T_u increases, the achievable energy of mMTC users decreases. By observing equations (4.18) and (4.26), it is clear that when the parameter T_u is relaxed, p_u^* decreases and when p_u^* decreases, p_m^* is reduced too. It has also already been proven that the mMTC energy consumption is an increasing function w.r.t. p_u^* and p_m^* . Thus, when transmission powers decrease, the objective function decreases too. From the same equations, it is straightforward that by increasing the task size L of devices, an increase in the transmission powers and thus in the energy consumption will occur too.

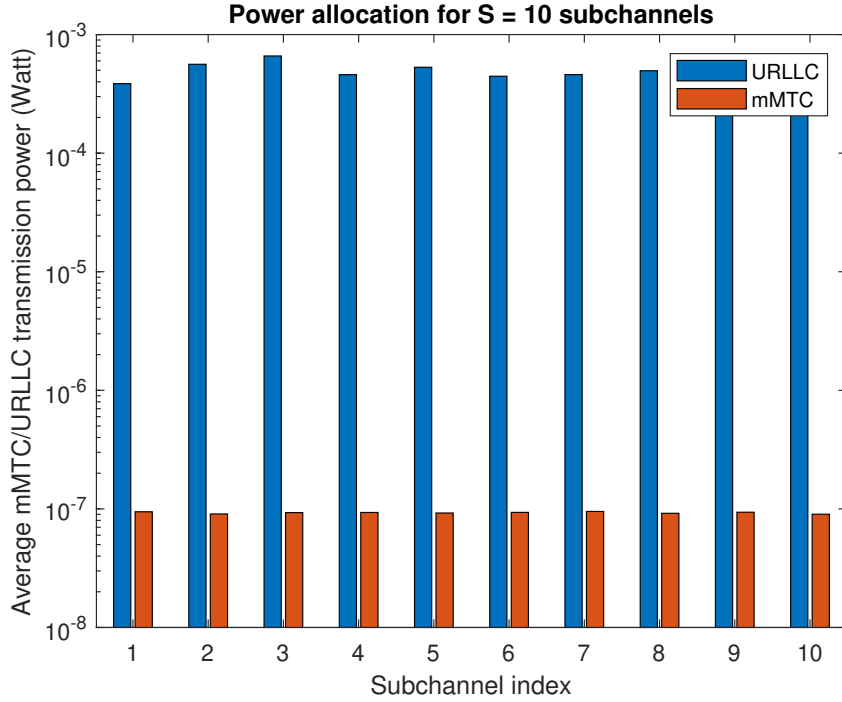


Figure 4.4: Average minimum power consumption of every mMTC and URLLC user in every subchannel for different channel gain coefficients ($E_u = 0.05\text{J}$, $T_m = 1.5\text{sec}$, $T_u = 0.03\text{sec}$)

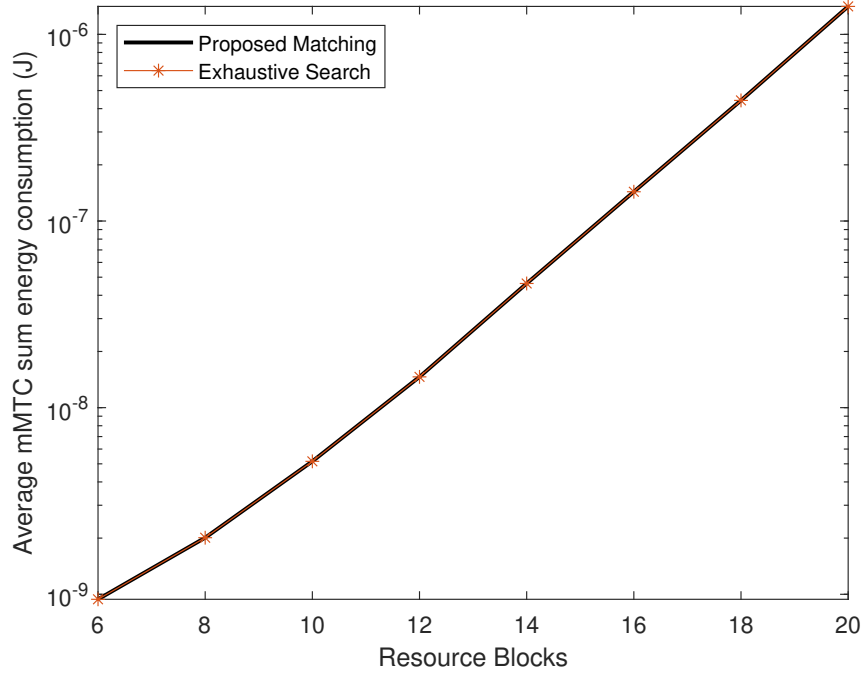


Figure 4.5: Average minimum sum energy consumption of mMTC users as a function of the resource blocks for different channel gain coefficients ($E_u = 0.05\text{J}$, $T_u = 0.03\text{sec}$, $T_m = 1.5\text{sec}$)

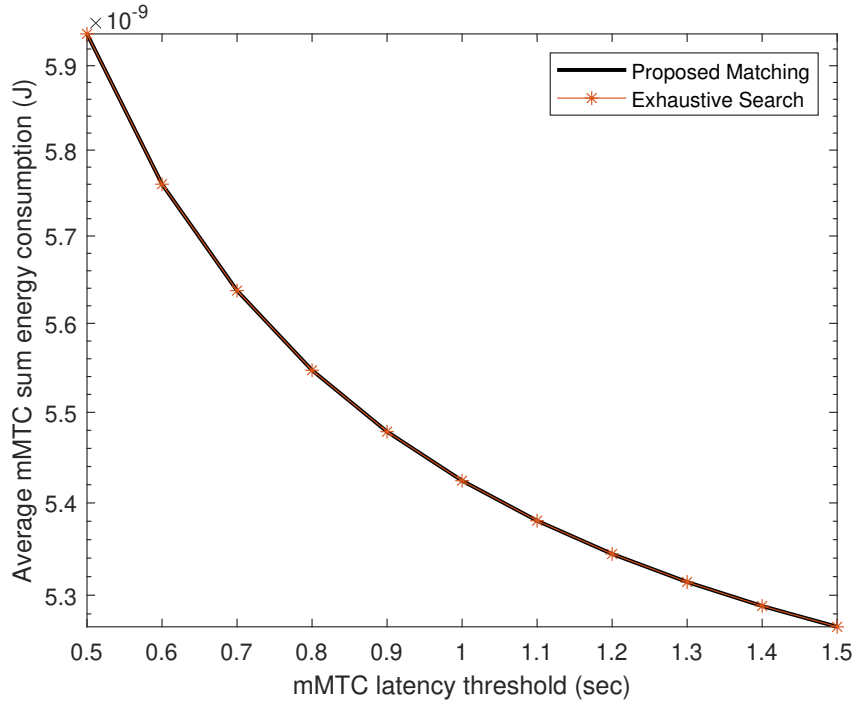


Figure 4.6: Average minimum sum energy consumption of mMTC users as a function of the latency threshold of URLLC users for different channel gain coefficients ($E_u = 0.05\text{J}$, $T_m = 1.5\text{sec}$)

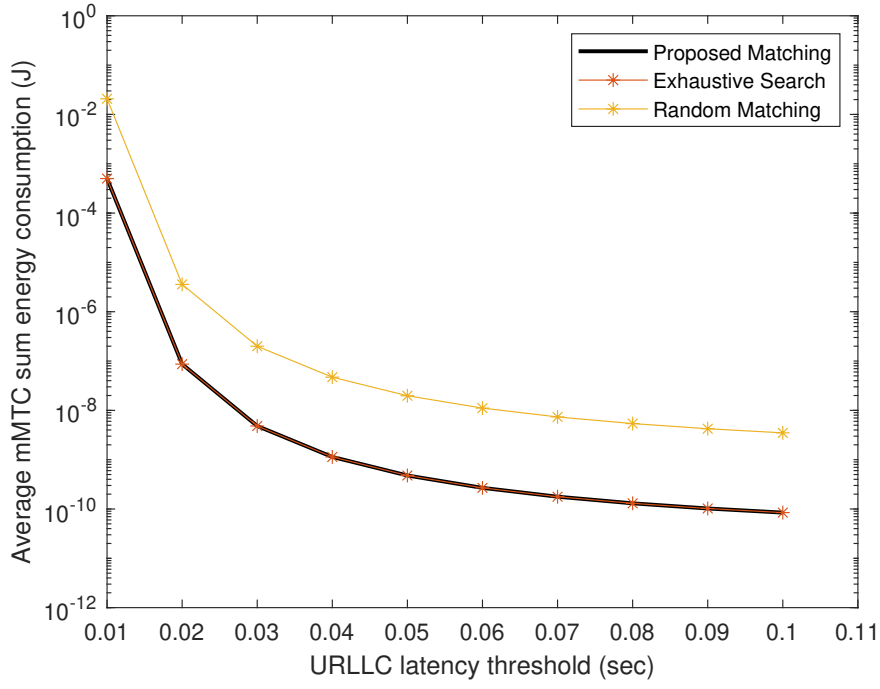


Figure 4.7: Average minimum sum energy consumption of mMTC users as a function of the latency threshold of mMTC users for different channel gain coefficients ($E_u = 0.05\text{J}$, $T_u = 0.03\text{sec}$)

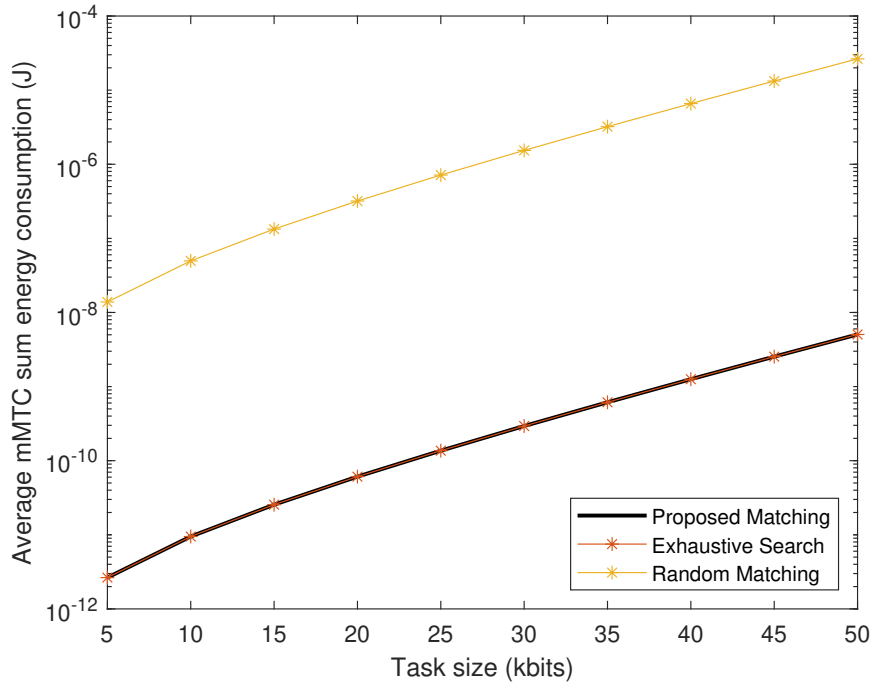


Figure 4.8: Average minimum sum energy consumption of mMTC users as a function of the task size of users for different channel gain coefficients ($E_u = 0.05\text{J}$, $T_u = 0.03\text{sec}$, $T_m = 1.5\text{sec}$)

Chapter 5

Conclusions

In this work, we have investigated the joint Mobile Edge Computing and the slicing of the Radio Access Network (RAN) resources to support the two out of three 5G traffic types, namely, URLLC and mMTC. We have initially considered an uplink Edge Computing scenario, in which we exploited the orthogonal approach in order to satisfy the resource requirements of both set of users. The solution of the proposed optimization problem was compared with selected benchmarks. The algorithms were repeated 1000 times for reliable and valid results. Our simulations have demonstrated that it is preferable to assume dynamic allocated bandwidth and CPU capacity in order to satisfy the stringent latency demands of URLLC users.

Moreover, the coexistence of mMTC and URLLC devices in a non-orthogonal manner has been studied in an uplink Mobile Edge Computing scenario. In this case, a problem decomposition strategy and an iterative algorithm were used to solve the formulated optimization problem. Once again, the algorithms were repeated 1000 times. Our simulations verified that the proposed methods can satisfy the stringent requirements of mMTC devices in low energy while keeping the latency demands of URLLC low. They also demonstrated the superiority of the low complexity Hungarian Algorithm solution compared to the random user selection approach.


Appendix A

Hungarian Algorithm Example

Figures A.1 to A.4. show an example of the pairing process based on the Hungarian method to minimize the energy consumption of mMTC users, with three mMTC users and three subchannels. The aforementioned steps in Algorithm 1 are conducted to find the optimal pairing, which is (m1, s2), (m2, s1) and (m3, s3).

Step 1 Row reduction: for each row, subtract the minimum value of the row from all the elements in that row

	s1	s2	s3
m1	4.44	2.62	2.31
m2	2.29	2.87	6.94
m3	3.15	2.72	2.17




	s1	s2	s3
m1	2.13	0.31	0
m2	0	0.58	4.65
m3	0.98	0.55	0

Figure A.1: Hungarian Method - Step 1

Step 2 Column reduction: for each column, subtract the minimum value of the column from all the elements in that column

	s1	s2	s3
m1	2.13	0.31	0
m2	0	0.58	4.65
m3	0.98	0.55	0



	s1	s2	s3
m1	2.13	0	0
m2	0	0.27	4.65
m3	0.98	0.24	0

Figure A.2: Hungarian Method - Step 2

Step 3 Cover all zeros in the rows and columns using a minimum number of horizontal or vertical lines

	s1	s2	s3
m1	2.13	0	0
m2	0	0.27	4.65
m3	0.98	0.24	0

→

	s1	s2	s3
m1	2.13	0	0
m2	0	0.27	4.65
m3	0.98	0.24	0

Figure A.3: Hungarian Method - Step 3

Step 4 Optimality check: (a) if the minimum number of covering lines is 3, an optimal assignment of zeros which don't lie in the same row or column is possible. The algorithm stops. (b) If the minimum number of covering lines is less than 3, then continue with Step 5.

	s1	s2	s3
m1	2.13	0	0
m2	0	0.27	4.65
m3	0.98	0.24	0

Algorithm stops since the covering lines

are 3.



Selected pair



Unselected pair

Figure A.4: Hungarian Method - Step 4

Bibliography

- [1] D. Soldani and A. Manzalini, “Horizon 2020 and beyond: on the 5G operating system for a true digital society,” *IEEE Vehicular Technology Magazine*, vol. 10, no. 1, p. 32–42, 2015.
- [2] A. Osseiran et al., “Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project,” *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, 2014.
- [3] R. ITU, *IMT Vision Framework and Overall Objectives of the Future Development of IMT for 2020 and beyond*. Rec. ITU-R M.2083, Sept. 2015.
- [4] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, “5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, p. 1201–221, 2017.
- [5] J. W. Won and J. M. Ahn, *3GPP URLLC patent analysis*. ICT Express, 2020.
- [6] M. Bennis, M. Debbah, and H. V. Poor, “Ultrareliable and low-latency wireless communication: Tail, risk, and scale,” *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.
- [7] “Study on scenarios and requirements for next generation access technologies (3gpp tr 38.913 version 14.2.0 release 14),” Sophia Antipolis, France, Tech. Rep., 03 2017.
- [8] “New services and applications with 5g ultra-reliable low latency communications,” 5G Americas, White Paper, 2018.
- [9] M. Siddiqi, X. Yu, and Joung, “5g ultra-reliable low-latency communication implementation challenges and operational issues with iot devices,” *Electronics*, vol. 8, p. 981, 09 2019.
- [10] R. Bhatia, B. Gupta, S. Benno, J. Esteban, D. Samardzija, M. Tavares, and T. V. Lakshman, “Massive machine type communications over 5g using lean protocols and edge proxies,” in *2018 IEEE 5G World Forum (5GWF)*, 2018, pp. 462–467.
- [11] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, “Massive machine-type communications in 5g: physical and mac-layer solutions,” *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.
- [12] “Network Slicing for 5G Networks and Services,” 5G Americas, White Paper.

- [13] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, “Network function virtualization: State-of-the-art and research challenges,” *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.
- [14] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, “Network function virtualization: Challenges and opportunities for innovations,” *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.
- [15] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, “Network function virtualization in 5g,” *IEEE Communications Magazine*, vol. 54, no. 4, pp. 84–91, 2016.
- [16] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, “A survey on software-defined networking,” *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 27–51, 2015.
- [17] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, “Software-defined networking: A comprehensive survey,” *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.
- [18] “SDN Architecture,” ONF TR-521, Tech. Rep., 02 2016.
- [19] “Applying SDN Architecture to 5G Slicing,” ONF TR-526, Tech. Rep., 04 2016.
- [20] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, “Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges,” *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.
- [21] A. Barakabitze, A. Ahmad, A. Hines, and R. Mijumbi, “5g network slicing using sdn and nfv: A survey of taxonomy, architectures and future challenges,” *Computer Networks*, vol. 167, p. 106984, 11 2019.
- [22] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [23] —, “A survey on mobile edge computing: The communication perspective,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [24] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, “Mobile Edge Computing A key technology towards 5G ,” ETSI, White Paper.
- [25] G. Brown, “Mobile Edge Computing Use Cases and Deployment Options ,” Juniper, White Paper.
- [26] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, “5g wireless network slicing for embb, urllc, and mmhc: A communication-theoretic view,” *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.
- [27] R. Kassab, O. Simeone, P. Popovski, and T. Islam, “Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures,” *IEEE Access*, vol. 7, pp. 13 035–13 049, 2019.

- [28] A. Anand, G. De Veciana, and S. Shakkottai, “Joint scheduling of urllc and embb traffic in 5g wireless networks,” in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1970–1978.
- [29] O. Vikhrova, C. Suraci, A. Tropeano, S. Pizzi, K. Samouylov, and G. Araniti, “Enhanced radio access procedure in sliced 5g networks,” in *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2019, pp. 1–6.
- [30] E. N. Tominaga, H. Alves, O. L. A. López, R. D. Souza, J. L. Rebelatto, and M. Latva-aho, “Network slicing for embb and mmhc with noma and space diversity reception,” in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, 2021, pp. 1–6.
- [31] A. E. Mostafa, Y. Zhou, and V. W. Wong, “Connectivity maximization for narrowband iot systems with noma,” in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.
- [32] L. Song, W. Qingming, S. Yanjing, and C. Xin, “Resource management for non-orthogonal multiple access based machine type communications,” 11 2018, pp. 413–419.
- [33] S. S. Metwaly, A. M. A. El-Haleem, and O. El-Ghandour, “Noma based matching game algorithm for narrowband internet of things (nb-iot) system,” *Ingénierie des Systèmes d’Information*, vol. 25, pp. 345–350, 06 2020.
- [34] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. Karagiannidis, “Optimal task assignment and power allocation for noma mobile-edge computing networks,” 04 2019.
- [35] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. K. Tsang, “Noma-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12 244–12 258, 2018.
- [36] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, “Delay minimization for noma-mec offloading,” *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1875–1879, 2018.
- [37] A. Kiani and N. Ansari, “Edge computing aware noma for 5g networks,” *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, 2018.
- [38] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, “Energy-efficient noma-based mobile edge computing offloading,” *IEEE Communications Letters*, vol. 23, no. 2, pp. 310–313, 2019.
- [39] M. Zeng and V. Fodor, “Energy-efficient resource allocation for noma-assisted mobile edge computing,” in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 1794–1799.
- [40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [41] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, “A tutorial on nonorthogonal multiple access for 5g and beyond,” *Wireless Communications and Mobile Computing*, vol. 2018, p. 1–24, Jun 2018. [Online]. Available: <http://dx.doi.org/10.1155/2018/9713450>

- [42] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. S. Kwak, “Non-orthogonal multiple access (NOMA): how it meets 5g and beyond,” *CoRR*, vol. abs/1907.10001, 2019. [Online]. Available: <http://arxiv.org/abs/1907.10001>
- [43] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109>