UNIVERSITY OF THESSALY

DOCTORAL THESIS

Design and Implementation of Resource Allocation Algorithms in 5G Heterogeneous Wireless Networks

Author:
Virgilios PASSAS

Supervisor:
Assoc. Prof. Athanasios KORAKIS

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Electrical and Computer Engineering

Dissertation Committee:
Assoc. Prof. Athanasios Korakis
Prof. Leandros Tassiulas
Assoc. Prof. Antonios Argyriou

March 2021

Declaration of Authorship

I, Virgilios PASSAS, declare that this thesis titled, "Design and Implementation of Resource Allocation Algorithms in 5G Heterogeneous Wireless Networks" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Περίληψη

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Σχεδιασμός και Υλοποίηση Αλγορίθμων Ανάθεσης Πόρων σε Ετερογενή Ασύρματα δίκτυα 5ης Γενιάς

Βιργίλιος Πασσάς

Τα επόμενης γενιάς (5G) και μετέπειτα δίκτυα κινητής τηλεφωνίας θα βασίζονται σε μεγάλο βαθμό σε κατανεμημένες ετερογενείς υποδομές, οι οποίες προβλέπεται να δημιουργήσουν ένα ευέλικτο και πολύπλευρο οικοσύστημα πόρων. Τα ετερογενή δίκτυα (HetNets) έχουν ήδη προταθεί ως λύση για το συνεχώς επιδεινούμενο πρόβλημα συμφόρησης των κυψελοειδών δικτύων. Οι τωρινές κινητές συσκευές είναι σε θέση να αξιοποιήσουν τα πλεονεκτήματα των HetNets, καθώς είναι εξοπλισμένα με πολλαπλές διεπαφές για ασύρματη πρόσβαση. Όμως το πρόβλημα επιλογής δικτύου σε ένα HetNet παραμένει ένα ανοιχτό θέμα προς έρευνα. Ταυτόχρονα, τα δίκτυα 5G επαναπροσδιορίζουν την αρχιτεκτονική του δικτύου, προτείνοντας τον διαχωρισμό της στοίβας πρωτοκόλλων του σταθμού βάσης και την υλοποίηση υπηρεσιών του δικτύου πρόσβασης χρηστών (RAN) μέσω λογισμικού. Αυτές οι προτάσεις βασίζονται στην έννοια του Cloud-RAN, όπου ένα κομμάτι του σταθμού βάσης εκτελείται ως λογισμικό σε υποδομές Νέφους (Cloud) και τα υπόλοιπα κομμάτια εκτελούνται στο hardware που συνδέεται ο πομποδέκτης. Στα Cloud-RANs υπάρχει η δυνατότητα να ενσωματωθούν διαφορετικές τεχνολογίες για πρόσβαση στο δίκτυο, προσφέροντας στους χρήστες περισσότερες επιλογές για την σύνδεση τους στο δίκτυο. Σε τέτοιες διατάξεις δικτύων, η αποτελεσματική κατανομή και διαχείριση της κίνησης δεδομένων είναι πρωταρχικής σημασίας. Μια άλλη σημαντική πρόοδος που φέρνει το 5G, είναι η ελαχιστοποίηση της καθυστέρησης πρόσβασης σε υπηρεσίες μέσω της εφαρμογής του Multi-access Edge Computing (MEC). Αυτή η λειτουργικότητα φέρνει τις υπηρεσίες στην άκρη του δικτύου, υποστηρίζοντας έτσι πολλές εφαρμογές που ανταλλάσσουν κρίσιμα χρονικά δεδομένα μέσω του δικτύου (π.χ. e-Health, Industry 4.0, AR / VR, κ.λπ.). Σε αυτή τη διατριβή, σχεδιάσαμε, μελετήσαμε και προτείναμε καινοτόμες λύσεις για την ανάθεση πόρων σε ετερογενή ασύρματα δίκτυα 5G με σταθμούς βάσης είτε υλοποιημένους ως μια οντότητα είτε κατανεμημένους, έχοντας την δυνατότητα παροχής MEC πόρων. Οι ερωτήσεις που προσπαθούμε να απαντήσουμε είναι οι εξής: 1) Πώς ένας χρήστης θα επιλέξει την καλύτερη τεχνολογία για πρόσβαση στο δίκτυο σε ένα ετερογενές σύστημα και πότε πρέπει να αλλάξει σε ένα άλλο; 2) Πώς ο πάροχος ενός τέτοιου ετερογενούς δικτύου πρέπει να χρεώνει την κάθε τεχνολογία πρόσβασης που διαθέτει; 3) Λαμβάνοντας υπόψη ένα ετερογενές Cloud-RAN, πώς μπορούμε να εξυπηρετήσουμε τους χρήστες με χαμηλή καθυστέρηση στην μετάδοση των πακέτων; 4) Πώς πρέπει να κατανέμονται οι MEC πόροι του δικτύου στους διαφορετικούς παρόχους υπηρεσιών; 5) Πώς θα χρεώνεται ένας πάροχος για τη μετακίνηση υπηρεσιών πιο κοντά στην άκρη του δικτύου με σκοπό την πρόσβαση των χρηστών σε υπηρεσίες με χαμηλή καθυστέρηση;

Αρχικά, ξεκινάμε με τις υπερ-πυκνές διατάξεις ετερογενών ασύρματων δικτύων. Η πολυπληθής και πυκνή ανάπτυξη δικτύων καθώς και η ενσωμάτωση ετερογενών τεχνολογιών

έχουν προταθεί ως λύση στον συνεχώς αυξανόμενο αριθμό συσκευών που συνδέονται στο Διαδίκτυο. Η ετερογένεια θεωρείται βασικό χαρακτηριστικό των δικτύων 5G καθώς επιτρέπει στους παρόχους κινητής τηλεφωνίας να μεταφέρουν χρήστες απο τα κυψελωτά δίκτυα τους σε παρακείμενα δίκτυα, μετριάζοντας έτσι την υπάρχουσα συμφόρηση και παρέχοντας υψηλής ποιότητας συνδεσιμότητα στους χρήστες. Ωστόσο, τέτοια περιβάλλοντα αυξάνουν την πολυπλοκότητα του προβλήματος ανάθεσης της δικτυακής κίνησης. Γι' αυτόν το σκοπό, σχεδιάζουμε και μελετάμε ένα ετερογενές σύστημα πρόσβασης στο δίκτυο που βασίζεται στο Paris Metro Pricing (PMP), λαμβάνοντας υπόψη τρεις διαφορετικές τεχνολογίες πρόσβασης (3G, 4G, WiFi). Παρουσιάζουμε ένα δυναμικό σχήμα τιμολόγησης για κάθε διαθέσιμη Radio Access Technology (RAT) του ετερογενούς δικτύου που λαμβάνει υπόψη την τρέχουσα διαθέσιμη χωρητικότητα κάθε RAT. Εξετάζουμε την απόδοσή του συστήματος όσον αφορά τη μέση τιμή της συνολικής ταχύτητας του συστήματος καθώς και τον αριθμό των εισερχόμενων χρηστών που μπορεί να δεχθεί, μελετώντας διάφορα σενάρια κινητικότητας. Αποτιμάμε το σύστημά μας χρησιμοποιώντας προσομοιώσεις και πειραματισμό σε πραγματικό περιβάλλον και παρουσιάζουμε τα αποτελέσματά μας.

Ως επόμενο βήμα, μελετάμε τις περιπτώσεις όπου οι χρήστες είναι multi-homed, δηλαδή διαθέτουν πολλαπλές διεπαφές πρόσβασης στο δίκτυο και μπορούν ταυτόχρονα να συνδεθούν με διαφορετικά RANs. Επιπλέον, θεωρούμε ότι οι χρήστες εκτελούν εφαρμογές με διαφορετικές απαιτήσεις δικτύου (χαμηλή καθυστέρηση μετάδοσης πακέτων, ταχύτητα κ.λπ.). Σε τέτοια σενάρια, το πρόβλημα ανάθεσης τεχνολογιών καθίσταται ακόμη πιο περίπλοκο, διότι η ποιότητα εμπειρίας (QoE) που αντιλαμβάνεται ο χρήστης εξαρτάται σε μεγάλο βαθμό από την εφαρμογή που εκτελεί. Για το σκοπό αυτό, προτείνουμε ένα σχήμα που προσπαθεί να αναθέσει αποτελεσματικά κάθε εφαρμογή στις διαθέσιμες τεχνολογίες του ετερογενούς συστήματος που μελετάμε με στόχο τη μεγιστοποίηση της συνάρτησης ωφέλειας του χρήστη. Μέσω της αποτίμησης της λύσης μας, επισημαίνουμε τη σημασία που έχει η κατάταξη των χρηστών κατά την διάρκεια ενημέρωσης του συστήματος. Για αυτόν τον λόγο, προτείνουμε και εξετάζουμε τρεις διαφορετικές πολιτικές κατάταξης των χρηστών (βάση των απαιτήσεων σε ρυθμό μετάδοσης των δεδομένων, της αποτελεσματικότητας του φάσματος και της ευαισθησίας των εφαρμογών). Τα αποτελέσματά μας δείχνουν πώς οι πολιτικές κατάταξης μπορούν να επηρεάσουν το κόστος που προκύπτει σε κάθε χρήστη, καθώς και το ποσοστό αξιοποίησης των διαθέσιμων RATs.

Στη συνέχεια, επικεντρωνόμαστε στο κατά πόσο είναι δίκαιες οι πολιτικές κατάταξης των χρηστών που έχουμε προτείνει στο σύστημα ανάθεσης πόρων που διερευνούμε σε 5G ετερογενή ασύρματα δίκτυα. Προς αυτήν την κατεύθυνση, παρουσιάζουμε μια βέλτιστη πολιτική κατάταξης των χρηστών που θα την χρησιμοποιήσουμε ως αναφορά και μέτρο σύγκρισης. Αυτή η πολιτική αξιοποιεί την έννοια της σταθμισμένης αναλογικότητας και τη συγκρίνουμε με τις τρεις προτεινόμενες πολιτικές κατάταξης λαμβάνοντας τη Kendall tau

συσχέτιση για καθεμία από αυτές. Με βάση τα ευρήματά μας, σχεδιάζουμε και εφαρμόζουμε ένα δυναμικό σχήμα επιλογής πολιτικής, το οποίο κάθε φορά που ενημερώνεται το σύστημα, επιλέγει μεταξύ των διαθέσιμων πολιτικών κατάταξης εκείνη που συσχετίζεται σε μεγαλύτερο βαθμό με την πολιτική αναφοράς. Εξετάζουμε περαιτέρω την τιμολόγηση του σχηματός μας, αρχικά, μόνο το διαθέσιμο εύρος ζώνης των τεχνολογιών πρόσβασης λήφθηκε υπόψη για το μοντέλο τιμολόγησης. Μέσω της αξιολόγησης της λύσης μας όμως, αποφασίσαμε να επεκτείνουμε το μοντέλο τιμολόγησης προκειμένου να συμπεριλάβουμε και τις απαιτήσεις των εφαρμογών που εκτελούνται στον χρήστη. Για την αξιολόγηση της λύσης μας, θεωρούμε 4 διαφορετικά RATs (3G, WiFi, 4G, WiGig / 5G-NR) για το 5G HetNet σύστημα που εξετάζουμε. Για την αποτίμηση του σχήματός μας, τόσο προσομοιώσεις όσο και πειραματισμός σε πραγματικές υποδομές χρησιμοποιούνται, επιτρέποντάς μας να εκτιμήσουμε την αποδοτικότητα κατανομής του εύρους ζώνης, τη διακύμανση των τιμών πρόσβασης, το πόσο δίκαιη είναι η κάθε πολιτική κατάταξης και το τελικό κόστος για τους χρήστες.

Επόμενος τομέας ενασχόλησής μας, είναι η μελέτη της μείωσης της καθυστέρησης για την πρόσβαση σε υπηρεσίες από τους χρήστες ενός ετερογενούς Cloud-RAN. Το ETSI έχει προτείνει το Multi-access Edge Computing (MEC) ως μέθοδο για την ελαχιστοποίηση της καθυστέρησης πρόσβασης σε υπηρεσίες, φέρνοντας τις υπηρεσίες πιο κοντά στον χρήστη, στην άκρη του ασύρματου δικτύου. Οι προδιαγραφές του 5G-New Radio (NR) καθορίζουν τον διαχωρισμό της στοίβας των πρωτοκόλλων των σταθμών βάσης με βάση το 3GPP Option-2 split που δημιουργεί δύο οντότητες: την Κεντρικοποιημένη Μονάδα (CU) και την Κατανεμημένη Μονάδα (DU) που παρέχει και το δίκτυο πρόσβασης. Η CU είναι σε θέση να διαχειριστεί πολλαπλές DU που μπορούν να είναι διαφορετικών τεχνολογιών πρόσβασης (5G-NR, 4G LTE ή ακόμη και WiFi). Προς αυτήν την κατεύθυνση, μελετάμε την τοποθέτηση υπηρεσιών MEC σε ένα Cloud-RAN με πολλαπλές ασύρματες τεχνολογίες. Ερευνούμε δύο διαφορετικές τοποθετήσεις των υπηρεσιών, είτε στο fronthaul κοντά με το DU είτε στο backhaul μαζί με το Core Network και για διαφορετικές τεχνολογίες πρόσβασης (LTE, WiFi). Τα πειράματά μας δείχνουν ότι η τοποθέτηση της υπηρεσίας κοντά στο σταθμό βάσης μειώνει την καθυστέρηση πρόσβασης σε υπηρεσίας ακόμη και με παλαιότερες τεχνολογίες (LTE, WiFi). Επιπλέον, διερευνούμε πώς επηρεάζεται η ποιότητα εμπειρίας (QoE) του χρήστη από την τοποθεσία της υπηρεσίας σε συνδυασμό με την ασύρματη τεχνολογία από την οποία θα εξυπηρετηθεί ο χρήστης.

Τέλος, μελετάμε την ανάθεση πόρων MEC σε 5G ετερογενή Cloud-RANs. Σε τέτοιες διατάξεις, χρήστες που διαθέτουν πολλαπλές ασύρματες διεπαφές έχουν την δυνατότητα να αποκτήσουν πρόσβαση σε MEC υπηρεσίες που βρίσκονται ακριβώς μετά το DU μέσω πολλαπλών ασύρματων συνδέσμων. Τα αποτελέσματα της έρευνάς μας, δείχνουν ότι η συγκεκριμένη τοποθέτηση των υπηρεσιών έχει καλύτερα αποτελέσματα ως προς την

καθυστέρηση σε σύγκριση με τις τοποθετήσεις των MEC υπηρεσιών που έχουν προταθεί απο το ETSI. Παρ’ όλα αυτά, η ασύρματη τεχνολογία μέσω της οποίας εξυπηρετείται ο κάθε χρήστης παίζει σημαντικό ρόλο στην καθυστέρηση και ποιότητα της εμπειρίας (QoE) που αντιλαμβάνεται ο χρήστης. Επιπλέον, οι τοποθεσίες για την τοποθέτηση των υπηρεσιών και η επιλογή των ασύρματων τεχνολογιών που χρησιμοποιούνται για την προώθηση των δεδομένων στους χρήστες μπορούν να αξιοποιηθούν ως παράμετροι διαφοροποίησης για τη χρέωση των MEC πόρων στους παρόχους υπηρεσιών. Με βάση αυτές τις διαπιστώσεις, σχεδιάζουμε και εφαρμόζουμε ένα μοντέλο για ανάθεση πόρων στους διάφορους υποστηριζόμενους συνδυασμούς τοποθέτησης MEC υπηρεσιών σε ένα Cloud-RAN με υποστηριζόμενες τεχνολογίες LTE και WiFi. Τα πειράματά μας δείχνουν ότι χρησιμοποιώντας πολλαπλά RATs ταυτόχρονα και με την κατάλληλη τοποθέτηση των MEC υπηρεσιών, η μέση καθυστέρηση μπορεί να διατηρηθεί κάτω από ένα όριο και παράλληλα ικανοποιεί τις απαιτήσεις της υπηρεσίας.

UNIVERSITY OF THESSALY

Abstract

Department of Electrical and Computer Engineering

Design and Implementation of Resource Allocation Algorithms in 5G Heterogeneous Wireless Networks

by Virgilios PASSAS

Next generation (5G) and beyond mobile networks will highly rely on distributed heterogeneous infrastructure, which is foreseen to create a versatile resource ecosystem. Heterogeneous Networks (HetNets) have already been proposed as a solution for the continuously deteriorating congestion problem of cellular infrastructures. Current mobile devices are able to utilize the HetNets since they are equipped with multiple radio access interfaces. Still the network selection problem in a HetNet is an open research issue. At the same time, 5G redefines the network architecture, through the introduction of base station disaggregation and the softwarization of the Radio Access Network (RAN). These features stem from the Cloud-RAN concept, where a part of the base station is running as a software in the Cloud and the rest on the hardware connected to the RF transceiver. These Cloud-RANs might aggregate multiple heterogeneous links on the access, allowing users to access the network with different technologies. In such network deployments, efficient traffic allocation and management is of prominent priority. Another important advancement of 5G, is the minimization of latency for accessing services through the application of the Multi-access Edge Computing (MEC). This functionality brings the services at the edge of the network, supporting several applications that exchange time-critical data over the network (e.g. e-Health, Industry 4.0, AR/VR, etc.). In this thesis, we design, study and propose innovative solutions for the resource allocation in 5G Heterogeneous Wireless Networks (with Base Stations either single unit or disaggregated) and with MEC capabilities. The fundamental questions that we try to answer are the following: 1) How should a UE choose the best Radio Access Technology (RAT) in a heterogeneous network and when should it change to a different RAT? 2) How should an operator of such a network charge each of the available RATs? 3) Given a heterogeneous Cloud-RAN, how can we serve end-users with low latency? 4) How should resources from a MEC enabled Cloud-RAN network be allocated to different Service Providers? 5) How would an operator be charged for moving services closer to the edge for providing lower latency access to the users?

Initially we begin with the ultra-dense deployments of heterogeneous wireless networks. Network densification and integration of heterogeneous technologies have been proposed as a solution to the constantly rising number of devices connected to the Internet. Heterogeneity is considered as a key feature of the 5G networks as it enables the mobile service providers to offload their cellular networks, thus mitigating the existing congestion and providing high quality connectivity to mobile users. Nevertheless, such environments add up to the complexity of the traffic allocation problem. To this aim, we design and study a heterogeneous network access scheme based on the Paris Metro Pricing (PMP), considering three different access technologies (3G, 4G, WiFi). We introduce a dynamic pricing scheme for each

available RAT of the heterogeneous network which takes into consideration the current available capacity of each RAT. We investigate its performance in terms of average throughput and acceptance capability of incoming users for several mobility scenarios, employing simulations and real-testbed experiments.

As our next step, we study the use case where the end-users are multi-homed, meaning that they feature multiple network interfaces and can be concurrently connected to different RANs. Moreover we consider that the end-users are running applications with different network demands (low latency, throughput, etc.). In such setups, the technology allocation problem becomes even more complicated because the perceived Quality of Experience (QoE) highly depends on the application that is running on the end-user. To this end, we propose a scheme that tries to efficiently map each application to the available technologies towards maximizing the end-user's utility function. Through the evaluation of our solution, we pinpoint the significance that the end-users' order has to the system during its updating phase. For this reason, we propose and examine three different ordering policies (based on data rate demands, spectrum efficiency, and sensitivity). Our results illustrate how the ordering policies may affect the cost incurred at each UE as well as the utilization of the available RATs.

Following this, we focus on the fairness of the UEs' ordering policies that we have introduced in our resource allocation scheme for 5G Heterogeneous Wireless Networks. Towards this direction, we introduce an optimal ordering policy as a reference. This policy leverages the weighted proportionality fairness concept. We compare it against the other proposed ordering policies taking the Kendall tau correlation for each one of the policies. Based on our findings, we design and implement a dynamic policy selection scheme, which selects among the available ordering policies the one which is closest to the reference policy every time that the system is updating. We further investigate the pricing of our scheme; initially, only the availability of access technologies' bandwidth was taken into consideration for our pricing model but through the evaluation of our solution, we decided to extend it in order to also include the rate demands of the applications running on the end-user side. For the evaluation of our solution, we consider 4 different RATs (3G, WiFi, 4G, WiGig/5G-NR) for our 5G HetNet. Both simulations and testbed experimentation are employed and enable us to assess the bandwidth allocation efficiency, the variation of access prices, the policy fairness and the induced cost to end-users.

The next part of our work is related to the minimization of service access latency. ETSI has proposed Multi-access Edge Computing (MEC) as the means to minimize the access latency by bringing the services closer to the wireless network edge. 5G-New Radio (NR) specifications define the disaggregation of the base stations based on the 3GPP Option-2 split which creates two entities: the Central Unit (CU) and the Distributed Unit (DU) which provides the actual access network. The CU is able to manage multiple DUs which can be of different access technologies (5G-NR, 4G LTE or even WiFi). Towards this direction, we study the placement of MEC services in a disaggregated network (Cloud-RAN) with multiple wireless technologies. We investigate two different deployments of the services, either on the fronthaul collocated with the DU or the backhaul collocated with the Core Network and for different access technologies (LTE, WiFi). Our testbed experiments denote that placing the service close to the base station reduces the service access latency even with legacy technologies (LTE, WiFi). Furthermore, we investigate how the UE's Quality

of Experience (QoE) is affected from the service placement in conjunction with the wireless technology that the UE will be served.

Finally, we study the resource allocation of MEC resources in 5G Heterogeneous Cloud-RANs. In such setups, multi-homed users are able to get access over multiple wireless links to services located just after the DU components of the cellular network, illustrating reduced network latency compared to conventional MEC deployments. Nevertheless, the technology through which each user may be served plays an important role in the overall perceived latency and Quality of Experience (QoE) of the end-user. Moreover, the locations for placement of the hosted services and wireless technologies used to forward data to the end-users can be exploited as differentiation parameters for charging application providers for hosting their services on the MEC platform. Therefore, we design and implement a system model for resource allocation across the different supported MEC combinations on a Cloud-RAN with LTE and WiFi technologies. Our testbed experiments highlight that utilizing multiple RATs at the same time and with the appropriate placement of the MEC service, the average latency can be kept below a threshold while meeting the service's requirements.

List of Publications

The results, ideas and figures of this thesis have been included in the following publications:

Journals

(J1) V. Passas, V. Miliotis, N. Makris, and T. Korakis. **"Pricing Based Distributed Traffic Allocation for 5G Heterogeneous Networks"**. In: *IEEE Transactions on Vehicular Technology Vol.69, No. 10, 2020, pp. 12111–12123*.

Conferences

(C1) V. Passas, N. Makris, C. Nanis, and T. Korakis. **"V2MEC: Low-Latency MEC for Vehicular Networks in 5G Disaggregated Architectures"**. *Accepted in 2021 IEEE Wireless Communications and Networking Conference (WCNC), 2021, IEEE*.

(C2) V. Passas, N. Makris, V. Miliotis, and T. Korakis. **"Pricing Based MEC Resource Allocation for 5G Heterogeneous Network Access"**. In: *2019 IEEE Global Communications Conference (GLOBECOM), 2019, IEEE*.

(C3) V. Passas, V. Miliotis, N. Makris, and T. Korakis. **"Dynamic RAT Selection and Pricing for Efficient Traffic Allocation in 5G HetNets"**. In: *2019 IEEE International Conference on Communications (ICC), 2019, IEEE*.

(C4) V. Passas, N. Makris, V. Miliotis, T. Korakis, and L. Tassiulas. **"MATCH: Multiple Access for Multiple Traffic Classes in 5G HetNets"**. In: *2018 IEEE International Conference on Communications (ICC), 2018, IEEE*.

(C5) V. Passas, V. Miliotis, N. Makris, T. Korakis, and L. Tassiulas. **"Paris Metro Pricing for 5G HetNets"**. In: *2016 IEEE Global Communications Conference (GLOBECOM), 2016, IEEE*.

Demonstrations

(D1) V. Passas, N. Makris, C. Nanis, and T. Korakis. **"MEC service placement over the Fronthaul of 5G Cloud-RANs"**. In: *2019 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN). 2019, IEEE*.

In addition, our research efforts within the same period led to the following publications that are not directly related to this thesis:

Journals and Magazines

(J1) V. Nejkovic, F. Jelekovic, N. Makris, V. Passas, T. Korakis, and M. Tosic. **"Semantic Coordination On the Edge of Heterogeneous Ultra Dense Networks"**. In: *Springer's Journal of Network and Systems Management Vol. 29, No. 17, 2021*.

(J2) V. Passas, N. Makris, Y. Wang, A. Apostolaras, A. Mpatziakas, A. Drosou,T. Korakis, and D. Tzovaras. **"Artificial Intelligence for Network Function Autoscaling in a Cloud-Native 5G Network"**. *Submitted to IEEE Network Magazine*.

Conferences

(C1) V. Miliotis, N. Makris, V. Passas, and T. Korakis. **"Portfolio Theory Application for 5G Heterogeneous Cloud-RAN Infrastructure"**. In: *2020 IEEE International Conference on Communications (ICC), 2020, IEEE*.

(C2) N. Makris, V. Passas, C. Nanis, and T. Korakis. **"On Minimizing Service Access Latency: Employing MEC on the Fronthaul of Heterogeneous 5G Architectures"**. In: *2019 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), 2019, IEEE*.

(C3) N. Makris, V. Passas, T. Korakis, and L. Tassiulas. **"Employing MEC in the Cloud-RAN: An Experimental Analysis"**. In: *2018 Workshop on Technologies for the Wireless Edge, 2018, ACM*.

(C4) N. Makris, A. D. Samaras, V. Passas, T. Korakis, and L. Tassiulas. **"Measuring LTE and WiFi coexistence in Unlicensed spectrum"**. In: *2017 European Conference on Networks and Communications (EuCNC), 2017, IEEE*.

Acknowledgements

This thesis represents a culmination of research work that has taken place at the Department of Electrical and Computer Engineering, University of Thessaly, Greece. Through this long endeavor, a group of people stood by me and it is certain that without their help, I would not have been able to complete my dissertation. I want to take a moment to thank them.

First and foremost, I would like to deeply thank my supervisor, Assoc. Prof. Thanasis Korakis for first taking me as an undergraduate and Master's student, guiding me through my first steps in research and supporting me throughout the years of my studies in University of Thessaly. I would like to express my sincere gratitude to Assoc. Prof. Thanasis Korakis and Prof. Leandros Tassiulas for giving me the opportunity to work in such an inspiring research team and to participate in European projects, broadening my knowledge in many aspects. I am also grateful to the members of my dissertation committee, Prof. Leandros Tassiulas, Assoc. Prof. Antonios Argyriou, Prof. Spyridon Lalis, Assoc. Prof. Gerasimos Potamianos, Assoc. Prof. Dimitrios Katsaros and Assoc. Prof. Iordanis Koutsopoulos for accepting to serve in the examination committe of my thesis.

I would also like to thank all my friends and labmates in NITLab for their support and collaboration everyday. Firts of all, I would like to thank Nikos Makris with whom I have been working and collaborating since the very beginning of my PhD thesis. Particular thanks go to Vasilis Miliotis for constructively assisting my work in all our joint research. I would really like to thank Ilias Syrigos and Kostas Chounos for the stimulating discussions and for all the fun we have had through all these years. Also the rest of the guys, Stratos Keranidis, Giannis Kazdaridis, Donatos Stavropoulos, Harris Niavis, Aris Dadoukis, Polychronis Symeonidis, Christos Zarafetas, Apostolis Apostolaras, Kostas Choumas, Alexandros Valantasis, Panagiotis Tzimotoudis, Antonis Kalkanof, Pavlos Basaras, Giannis Zografopoulos, Panagiotis Skrimponis, Kostas Katsalis, Giannis Igoumenos, Dimitris Giatsios, Panagiotis Karamichailidis, Christos Nanis, Giannis Mavridis, Giorgos Theodosiou, Christos Theologou and Vasilis Zafeiris. I would also like to thank Christina Madelou and Katerina Arvaniti for their daily support in administration related issues. My sincere thanks to my closest friends, Christos, Thanasis, Mpampis, Kostas, Giorgos and Giannis for their encouragement all these years.

Last but not least, I want to thank my family, my wife Foteini, my parents Annoula and Grigoris, my grandmother Stella, my aunt Katerina and the ones that have passed away (grandparents Giannis, Vergos, Eleni and my uncle Dimitris). I have no words to express my gratitude for them, their support, patience and unconditional love for all my decisions. The least I can do in recognition to their love and support is to dedicate this dissertation to them.

Dedicated to my family.

Contents

List of Figures

List of Tables

List of Abbreviations

3GPP	**3**rd **G**eneration **P**artnership **P**roject
AP	**A**ccess **P**oint
AR	**A**ugmented **R**eality
BS	**B**ase **S**tation
CAPEX	**C**apital **E**xpenditure
CN	**C**ore **N**etwork
CNC	**C**entralized **N**etwork **C**ontroller
COTS	**C**ommercial **O**ff-**T**he-**S**helf
CU	**C**entralized **U**nit
DASH	**D**ynamic **A**daptive **S**treaming over **H**TTP
DL	**D**ownLink
DU	**D**istributed **U**nit
eNB	**e**volved **N**ode**B**
EPC	**E**volved **P**acket **C**ore
ETSI	**E**uropean **T**elecommunication **S**tandards **I**nstitute
FH	**F**ront**H**aul
HetNet	**Het**erogeneous **Net**work
HSPA	**H**igh **S**peed **P**acket **A**ccess
IP	**I**nternet **P**rotocol
LTE	**L**ong **T**erm **E**volution
LWAAP	**L**TE **W**LAN **A**ggregation **A**daptation **P**rotocol
LXC	**L**inu**X** **C**ontainer
MAC	**M**edium **A**ccess **C**ontrol
MCS	**M**odulation and **C**oding **S**cheme
MEC	**M**ulti-access **E**dge **C**omputing
MIMO	**M**ultiple **I**nput **M**ultiple **O**utput
MNO	**M**obile **N**etwork **O**perator
mmWave	**m**illi**m**eter **Wave**
MPD	**M**edia **P**resentation **D**escription
NITOS	**N**etwork **I**mplementation **T**estbed using **O**pen **S**ource platforms
NR	**N**ew **R**adio
OAI	**O**pen**A**ir**I**nterface
OPEX	**O**perational **E**xpenditure
OSI	**O**pen **S**ystems **I**nterconnection
PDCP	**P**acket **D**ata **C**onvergence **P**rotocol
PMP	**P**aris **M**etro **P**ricing
QoE	**Q**uality **o**f **E**xperience
QoS	**Q**uality **o**f **S**ervice
RAN	**R**adio **A**ccess **N**etwork
RAT	**R**adio **A**ccess **T**echnology
RLC	**R**adio **L**ink **C**ontrol
RNTI	**R**adio **N**etwork **T**emporary **I**dentifier
RTT	**R**ound **T**rip **T**ime

SAP	**S**ervice **A**ccess **P**oint
SCTP	**S**tream **C**ontrol **T**ransmission **P**rotocol
SDR	**S**oftware **D**efined **R**adio
SP	**S**ervice **P**rovider
TC	**T**raffic **C**lass
TCP	**T**ransmission **C**ontrol **P**rotocol
UDP	**U**nreliable **D**atagram **P**rotocol
UE	**U**ser **E**quipment
UHD	**U**ltra **H**igh **D**efinition
UL	**U**plink**L**ink
UMTS	**U**niversal **M**obile **T**elecommunications **S**ystem
USRP	**U**niversal **S**oftware **R**adio **P**eripheral
VNF	**V**irtual **N**etwork **F**unction
VR	**V**irtual **R**eality
WiFi	**Wi**reless **Fi**delity
WiGig	**Wi**reless **Gig**abit Alliance
WLAN	**W**ireless **LAN**

Chapter 1

Introduction

Contents

1.1 Motivation and Problem Statement

The fifth generation of mobile networking is fostering several advancements in the access, edge and core network, promising to offer higher network capacity with lower latency, allowing a variety of (critical) services to thrive around this ecosystem. One of the major enablers for 5G mobile networks is the heterogeneity in the Radio Access Technologies (RATs). The Ultra-dense heterogeneous networks (UDHetNets) are expected to play a key role not only during the transition from legacy (4G) to next-generation (5G) mobile networks but also when the 5G networks will be fully functional and established. Due to spectrum crunch in the sub-6GHz band, the dense-deployment of multiple access networks operating in different frequencies, offers to the operators extra capacity of either smaller scale access points (e.g. femto-/pico-cells) with existing or legacy technologies (e.g. LTE, HSPA, UMTS) and non-3GPP technologies like WiFi, complementing the macro-cell base station, towards offering enhanced capacity and lower latency services to the connected mobile network users.

At the same time, through the utilization of new wireless spectrum, available at the centimeter and millimeter Wave (mmWave) bands for fixed wireless access, ultra-high speed broadband services can be offered. The above technologies are widely considered as the main enablers for the 5th generation of mobile networking, annotated as 5G communications. Nevertheless, these technologies induce increased costs for the operators: from the mobile network operator's (MNO) perspective, induced CAPEX and OPEX costs on upgrading and maintaining the network respectively are higher, whereas the Average Revenue per User (ARPU) is slowly decreasing. Therefore, solutions that do not disrupt the cost structure of operators need to be applied towards keeping up with the constantly rising user demands and the cellular network evolution, thus the Heterogeneous Networks (HetNets) are considered as one of them.

On the other side, mobile devices nowdays are equipped with a plethora of network interfaces and with the introduction of multi-homing protocols (MPTCP,

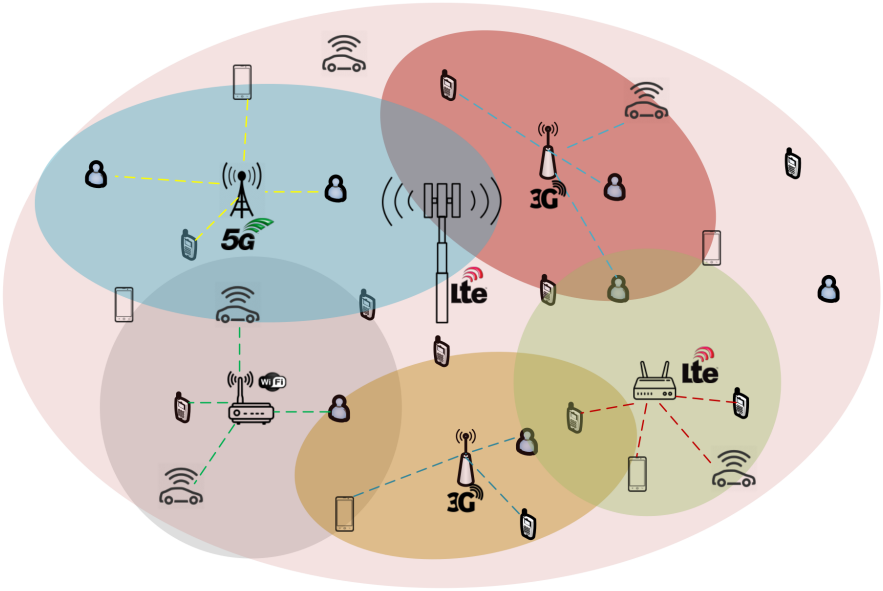


FIGURE 1.1: Ultra-dense heterogeneous network environment

SCTP), the exploitation of multiple paths over multiple networking interfaces associated with different RATs is now a reality. Considering a HetNet with multi-homed users, several issues are emerged. From the network operator perspective, the increased heterogeneity in the RATs, the diversity in the coverage ranges and the cappacities of each cell, create a highly complex environment for selecting the appropriate combination of technologies per UE. The multi-homing features only add up to the complexity of maximizing the overall performance of the HetNet. From the UE perspective, the different access technologies may induce different charges for the end-user thus the selection among the available RATs should take also into consideration the cost per access technology.

Another key feature of 5G networks will be the Cloud-based Radio Access Network (Cloud-RAN) where the base station changes from a single monolithic unit to a Baseband Unit (BBU) and a Remote Radio Unit (RRU). The RRU can be considered either as passive, with sole purpose to transmit over the air tha data that is received, or intelligent, with parts of the processes of the stack taking place on it. Towards the direction of an intelligent RRU, selecting the point of the stack that the split is applied is of high importance. The higher layer the split takes place, the lower the fronthaul requirements are, thus reducing the burden placed on the fronthaul interface. Furthermore, there are points for disaggregating the base station where the integration of heterogeneous RRUs in the system is applicable.

5G also benefits from the wide application of Multi-access Edge Computing (MEC), by serving network users directly from the network edge. Employing MEC can drastically reduce service-to-UE latency and ETSI has specified the different deployments for MEC services (Fig. 1.2) as follows: 1) the bump-in-the-wire mode, where the service is located just after the base station, relieving the network from the extra delay added for sending traffic to the Core Network, 2) the case of collocating the MEC services with the Core Network at an Edge datacenter, which has the benefit of handling IP traffic just after the Core Network, and 3) the local break-out mode where a part of the Core Network is handling only data plane traffic collocated with the base station, whereas the control plane traffic is sent to a traditional Core Network deployment.

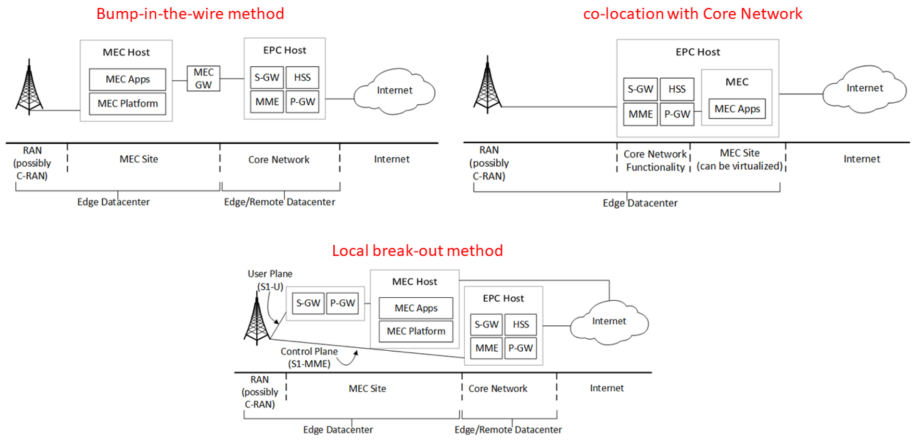


FIGURE 1.2: Existing MEC placement methods defined by ETSI

The different suggested placements for the MEC service provide fertile ground for the differentiation of the hosted service providers and how the infrastructure providers will offer their MEC resources. Moreover, considering these MEC setups to be employed in a heterogeneous Cloud-RAN more open research issues arise in terms of resource allocation of MEC services, RAT and their possible combinations. Another open issue related to MEC, is that its integration considers the same placements for the edge resources as in previous solutions such as the LTE despite the new architecture of Cloud-RAN and 5G-NR. This observation paves the way for the development of new solutions that are mapped to the new (5G) architecture.

Goal of this thesis is to address these open research issues presented above and are summarized as follows : 1) In a Multi-RAT system, which is the best access network for every client of the system to connect with, 2) Supporting multi-homed clients in a Heterogeneous Network make the resource allocation problem more complex, 3) Current MEC placements merely mapped to the new 5G network architecture as its integration considers the same setups as in the legacy networks, 4) Utilization of MEC resources in heterogeneous Cloud-RANs adds up to the complexity of the resource allocation in such environments.

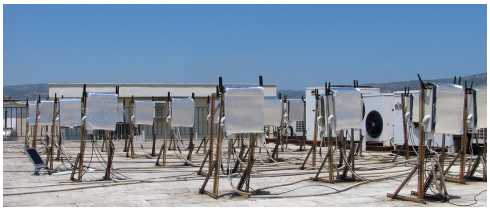
1.2 Experimental Tools and Methods

In this section, we describe the main tools that we use for evaluating the proposed solutions in this thesis. These are broken down in the following: 1) NITOS Facility, in our experiments we use the NITOS wireless testbed which is located in University of Thessaly, Greece and 2) the OpenAirInterface platform, used to form real cellular base stations over commodity hardware. Below we present details for these two components.

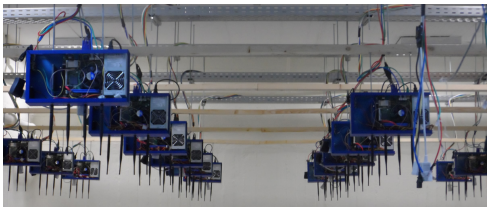
1.2.1 NITOS Facility

NITOS Facility is an integrated facility with heterogeneous testbeds that focuses on supporting experimentation-based research in the area of wired and wireless networks. NITOS is remotely accessible and open to the research community 24/7 through the NITOS portal, allowing users from around the globe to take advantage of highly programmable equipment. The testbed is based on open-source software

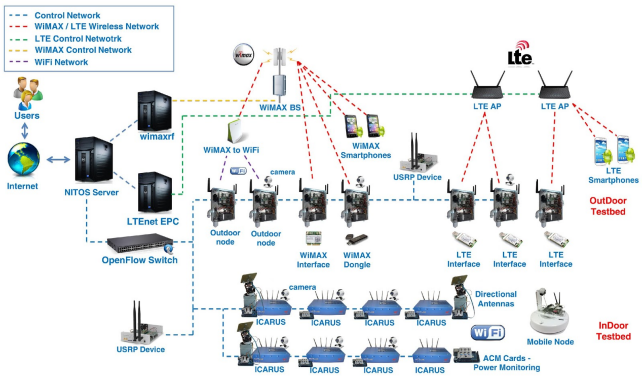
that allows the design and implementation of new algorithms, enabling new functionalities on the existing hardware. Parallel experimentation (slicing) of different users is enabled, through the utilization of the NITOS scheduler software. NITOS has an established user base of over 4000 users in the past years, with over 20 researchers using the infrastructure in a daily basis. It is federated with several infrastructures all over the world (Europe, Brazil, South Korea) in the context of various projects, like OpenLab, Fed4FIRE, SmartFIRE while it is also part of the OneLab federation. Services currently offered by the infrastructure are the following:



(A) Outdoor testbed



(B) Indoor RF isolated testbed



(C) NITOS architecture

FIGURE 1.3: NITOS Facility Testbeds

- A wireless experimentation testbed, which consists of 100 powerful nodes (some of them mobile) in indoor and outdoor deployments that feature multiple wireless interfaces and allow for experimentation with heterogeneous (Wi-Fi, WiMAX, LTE, Bluetooth) wireless technologies.

- A wireless sensor network, consisting of a controllable testbed deployed in an indoor environment, a city-scale sensor network in Volos city and a city-scale mobile sensing infrastructure that relies on bicycles of volunteer users. Most of the sensor platforms are custom-made, developed by UTH, and some others commercial, all supporting open-source and easy to use firmware and exploit several wireless technologies for communication (ZigBee, Wi-Fi, BLE, LoRa and 6LoWPAN).

- A Software Defined Radio (SDR) testbed that consists of multiple USRP devices attached to the NITOS wireless nodes. USRPs allow the researcher to program a number of physical layer features (e.g. modulation), thereby enabling dedicated PHY layer or cross-layer research.

- A mmWave testbed, operating in the V-band (60GHz), consisting of six different nodes supporting multi-Gbps over the air speeds, and beam-steering with 15 degrees step.

- A drone base testbed, consisting of five high-performing drones that are able to carry NITOS nodes and setup wireless mesh setups with different technologies (WiFi, mmWave).
- A Software Defined Networking (SDN) testbed that consists of multiple OpenFlow technology enabled switches, connected to the NITOS nodes, thus enabling experimentation with switching and routing networking protocols.
- A Cloud infrastructure, which consists of 7 HP blade servers and 2 rack-mounted ones providing 272 CPU cores, 800 GB of Ram and 22TB of storage capacity, in total. For the provisioning of the cloud, OpenStack is used.

The different testbeds of NITOS are interconnected, thus supporting a wide range of experiments, from simple one to more complex, utilizing heterogeneous resources from all the NITOS testbeds.

1.2.2 The OpenAirInterface Platform

OpenAirInterface (OAI) wireless technology platform is the first open-source software-based implementation of the LTE system spanning the full protocol stack of 3GPP standards. It features contributions both in E-UTRAN (wireless access - eNB & UE) and the Evolved Packet Core (EPC). It can be used to build and customize an LTE base station and core network on a PC and connect commercial UEs to test different configurations and monitor the network and the mobile devices in realtime. OAI is based on a PC hosted software radio frontend architecture where the transceiver functionality is realized via an SDR device connected to a host computer for processing. OpenAirInterface provides both UE, eNB, and core-network functionality. OAI is written in standard C and is released as free software under the terms of version 3 of the GNU General Public License (GPLv3). In this thesis, we utilize the OAI platform for the experimental evalaution of our proposed solutions. We use OAI either out-of-the-box or extending its functionalities by adding new features towards the implementation of MEC services close to the Base Station.

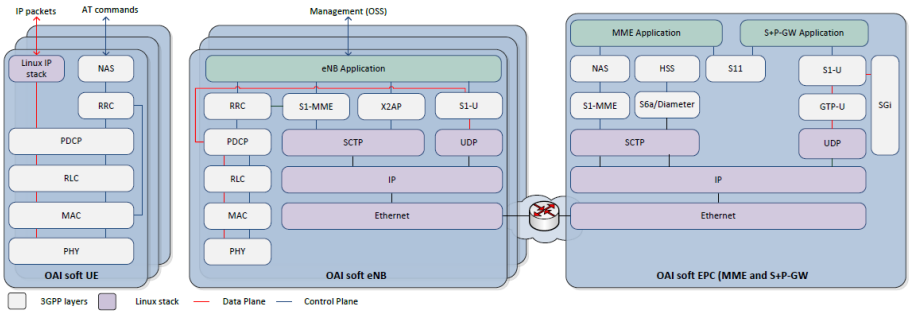


FIGURE 1.4: OpenAirInterface protocol stack

1.3 Thesis Synopsis

In this thesis, we study resource allocation algorithms in 5G Heterogeneous networks and how to minimize the service access latency on such architectures by employing the MEC concept. The common research approach that has been followed across all of our contributions in this thesis, is based on the design and implementation of the proposed algorithms in real commodity and experimental hardware.

This approach enabled us to assess the performance of the developed schemes under realistic environments. The fundamental questions that we try to answer are the following:

1. How a UE should choose the best RAT in a heterogeneous network and when it should change to a different RAT?
2. How should an operator of such network charge each of the available RATs?
3. Given a heterogeneous Cloud-RAN, how can we serve end-users with low latency?
4. How should resources from a MEC enbaled Cloud-RAN network be allocated to different Service Providers?
5. How would an operator be charged for moving services closer to the edge for providing lower latency access to the users?

In the following paragraphs, we provide a summary of the works included in this thesis, towards addressing these questions.

Initially, in **Chapter 2** we focus on the heterogeneous ultra-dense networks and the resource allocation algorithms that could be applied in such environments. Open issues in such deployments are for example; how an end-user should choose the best of the available Radio Access Technologies (RATs) or when it should change RAT. Towards this direction, in this chapter we propose a heterogeneous network access solution based on the Paris Metro Pricing (PMP), a service differentiation scheme that was first used in Paris metro to give to its passengers the ability to opt for less congested wagons. We extend the PMP policy by introducing a dynamic pricing scheme for each available RAT and we investigate its performance under a realistic mobility model for users in an urban environment. Our contributions are a step towards the direction of future 5G deployments, which are envisioned with dense multi-RAT heterogeneous networks, creating the challenge for the development of mechanisms that efficiently aggregate the capacity and coverage of diverse existing access technologies.

Chapter 3 presents our contributions towards extending the aforementioned resource allocation algorithm by enabling an end-user to be concurrently connected to multiple RATs. The technology allocation problem becomes even more complicated when considering that the perceived Quality of Experience (QoE) highly depends on the application that is running on the end-user. Different applications have versatile demands for the network capacity; thus a user with multiple applications might benefit from having simultaneously several connections of different technologies active to the Internet. Our proposed scheme tries to efficiently map each application to the available technologies towards maximizing the client's utility function. Evaluating our solution, we infered that during the update period of the system, the UE's ordering plays a significant role on the UE's incurred cost as well as the utilization of the RATs on the network. Furthermore, we propose three different ordering policies (based on data rate demands, spectrum efficiency, and sensitivity) for the UEs and present experimental results obtained from the application of our proposed model in a real testbed environment.

Following this, in **Chapter 4** we focus on the fairness of the UEs' ordering policies introduced in our previous contributions. The results of the aforementioned scheme

pinpoint the importance of the UE ordering when the system is updating and how it affects the UE's cost and the utilization of the system's RATs. For this purpose, we introduce an optimal ordering policy as a reference which leverages the weighted proportionality fairness concept and we compare it against the other proposed ordering policies through the Kendall tau correlation. Given the obtained results of this comparison, we propose a dynamic policy selection scheme that in each system cycle, the network controller which is responsible for the operation of the system, selects among the available ordering policies, the one which is closest to the reference policy. Furthermore, a second pricing for the radio access is proposed which is based not only the availability of access technologies' bandwidth but also the rate demands of the UEs' traffic classes. Both simulations and testbed experimentation are employed for the evaluation of the proposed solution and give us insights on the policy fairness, the variation of access prices, the bandwidth allocation efficiency and the induced cost to UEs.

In **Chapter 5**, we study the minimization of service access latency on a Heterogeneous Cloud-RAN. Towadrs this direction, ETSI proposed Multi-access Edge Computing (MEC) as the means to minimize the access latency by bringing the services closer to the wireless network edge. At the same time, Ultra-dense heterogeneous networks are expected to play a key role during the transition to 5G netowrks. In this chapter, we design and implement different placements of the MEC service (either on the fronthaul or the backhaul). Moreover, we study if the access technology impacts the service access latency, employing two different access technologies (LTE, WiFi). We validate our contributions in a real-world environment and present our findings. Our experimental results give insights on how the UE's Quality of Experience (QoE) is affected from the service placement in conjuction with the wireless technology that it will be served.

Finally, in **Chapter 6** we design and implement a scheme for the resource allocation of MEC resources to different service providers. Two pricing policies are proposed and applied in our design. Our scheme takes into consideration heterogeneous Cloud-RANs with MEC resources available in different places of the network (fronthaul, backhaul). Knowing that different RATs don't achieve the same latency times, enable us to propose different pricing for each service per each end-user. Furthermore, having the deployment of a heterogeneous Cloud-RAN with MEC capabilities network in the NITOS testbed, we are able to evaluate the proposed scheme under a real-world environment and not only through simulations. Our extensive experiments highlight the efficiency of our scheme to keep the average latency below a threshold while meeting the requirements of each MEC service of the system.

Chapter 2

Employing Paris Metro Pricing in 5G HetNets

Contents

2.1 Chapter Introduction

In recent years, global mobile data demand presents a continuous growth. According to Cisco Visual Networking Index [25], the number of mobile devices is also expected to present a vast increase, as any device from anywhere is already able to access the internet through cellular infrastructures, following the concept of Internet of Everything (IoE). These trends are already putting pressure on the mobile service providers, as the update of existing infrastructures cannot follow the growth rate of mobile data demand. Millimeter wave (mmW) radio access will increase the available spectrum for mobile devices in future 5G networks [36]. Nevertheless, until the deployment of the 5G technology infrastructure takes place, the use of multiple Radio Access Technologies (RAT) has been proposed, aiming to mitigate the congestion conditions already met by the mobile service providers. Heterogeneity will be a key feature of 5G networks [111], giving to providers the ability to offload their cellular networks to mitigate congestion and provide high quality connectivity to mobile devices.

Contemporary mobile devices are equipped with multiple radio access capabilities. At the same time multiple radio access technologies are deployed by mobile service providers (WiFi, 3G, LTE, small-cells etc), mainly in urban areas, giving to customers the option to choose and connect to different access networks.

The main questions that we are aiming to tackle in this chapter are:

1. How a User Equipment (UE) should choose the best of the available RATs.
2. When it should change to a different RAT.
3. How changes of access decisions affect the performance of other users in the multi-RAT environment.

Towards this direction, we propose a heterogeneous network access solution based on the Paris Metro Pricing (PMP), a service differentiation scheme that was first used in Paris metro to give to its passengers the ability to opt for less congested wagons. In this work, we extend the PMP policy by introducing a dynamic pricing scheme for each available service class and we investigate its performance under a realistic mobility model for users in an urban environment, and under real testbed experiments.

The rest of this chapter is organised as follows. In Section 2.2 we present a literature overview of heterogeneous network access schemes. In Section 2.3 we present the system model of the dynamic PMP scheme. In Section 2.4 we provide our proposed algorithms and the system architecture. Section 2.5 includes our evaluation results on the performance of the proposed scheme in terms of average throughput and acceptance capability of incoming users. Finally, Section 2.6 concludes our work.

2.2 Related Work

Many research groups have recently proposed different access schemes to exploit the coexistence of multiple access technologies, aiming to improve the Quality of Service (QoS) of mobile users and to mitigate the network congestion. In [112], the authors provide an extensive classification of published research works on network selection for heterogeneous networks, mainly regarding the utility functions and the mathematical models that were used. A survey on key parameters for handover decisions in heterogeneous networks is presented in [114]. A reward based algorithm is presented in [27], where mobile users autonomously update the fraction of their traffic through each available access technology, based on the rewards received by the base stations. These rewards are sent to each user, representing the impact of their traffic on the cell load. The authors in [8] formulate the multi-user RAT selection problem as a non-cooperative game, where each user tries to selfishly maximize its own throughput, and they investigate the impact of a user's decisions on other users performance and the convergence of the system to Nash equilibria. An incentive mechanism that aims to motivate WiFi Access Points (APs) to participate in heterogeneous networks, by providing an access class to the existing cellular infrastructure, is proposed in [58]. The pricing strategy for the inclusion of third party WiFi APs is formulated as a Stackelberg game between the mobile network provider and the third party WiFi APs.

Another model that creates on-demand multi-RAT conditions is proposed in [57], where mobile users form short range mesh networks to collaborate, by sharing their internet access with provision for proper routing policies with load-balancing and fairness. The authors also propose a virtual currency to create incentives and facilitate the cooperation of users. A pricing-based proportionally fair scheme for concurrent uplink access through LTE and WiFi is proposed in [75]. The seminal work for the application of PMP in the context of packet delivery networks was provided in [85] where the use of PMP was presented as a solution to the congestion control

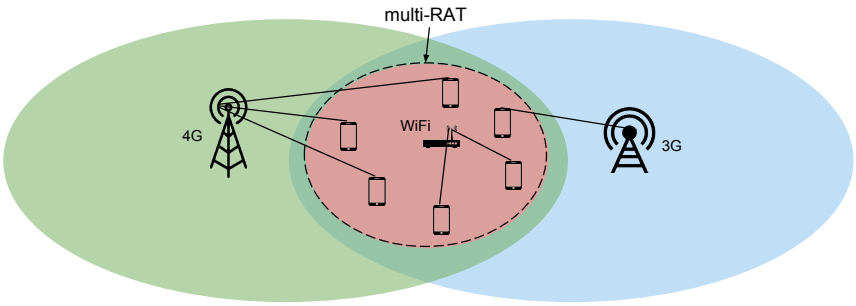


FIGURE 2.1: Multi-Radio Access Technology Environment.

problem for differentiated classes of service levels. The authors in [18] consider a general model of congestion externality for the PMP and investigate sufficient conditions of congestion functions that guarantee the viability of the PMP scheme. Inspired by these two works, we propose an extended PMP scheme with dynamic pricing, where users entering a multi-RAT environment decide which technology they prefer to access, depending on its current price and their sensitivity to congestion. In [51], the authors focus on pricing different classes of services from the service providers perspective and show that if prices are not chosen properly, matching the service qualities of the available classes, the system may settle into an undesirable equilibrium similar to the Prisoner's Dilemma game. Furthermore, they show that dynamic pricing based on users' preferences over different service classes can lead to stable equilibrium. This work is a step towards the direction of future 5G deployments, which are envisioned with dense multi-RAT heterogeneous networks [109], creating the challenge for the development of mechanisms that efficiently aggregate the capacity and coverage of diverse existing access technologies.

2.3 PMP System Model

We consider the existence of M classes of radio access services that belong to the same cellular network provider. Every access class m corresponds to a radio access technology (e.g. 4G, 3G, WiFi) and has a capacity C_m. Hence, the total system capacity is equal to $C = \sum_{m=1}^{M} C_m$. We focus on UEs that are under the coverage of all provided radio access technologies as depicted in Figure 2.1. Every UE is characterized by its type $\theta_i \in [0,1]$, representing its sensitivity to congestion. A UE_i with congestion sensitivity θ_i, using access class m has a utility equal to:

$$U_i(m) = V - p_m - \theta_i f(Q_m, \bar{C_m}) \tag{2.1}$$

where V is a flat-rate valuation of accessing the multi-RAT service, p_m is the access charge for a UE served at class m and $f(Q_m, \bar{C_m})$ is a function for the perceived congestion at class m. The mass of users choosing class m is denoted by Q_m, and the available capacity of the access class m is denoted by $\bar{C_m}$. Without loss of generality, we assume that there is a total mass of UEs equal to 1. The mass of users that do not participate in the proposed service is equal to $Q_0 = 1 - \sum_{m=1}^{M} Q_m$. We also assume that $V \geq p_1 > p_2 > ... > p_N$ and therefore a UE_i of type θ_i will choose the service class that maximizes its utility:

$$m(\theta_i) = \underset{1 \leq m \leq M}{\text{argmax}}\, U_i(m) \tag{2.2}$$

or no service if $U_i(m) = 0, \forall m \in (1, ..., M)$. This leads to a two-level (Stackelberg) game, where the provider first decides the prices per class, which are a function of the already allocated capacity per access technology, and then the UEs distribute themselves over classes, selecting the most appropriate for them. The provider can play by anticipating the distribution of UEs over the provided service classes. The distribution of UEs over the provided access technologies will be a Wardrop equilibrium [113], at which no UE will have an interest in changing access classes. When equilibrium of distribution of UEs over available access classes is reached, a given UE of type θ_i will prefer class m over k if

$$V - p_m - \theta_i f(Q_m, \bar{C_m}) \geq V - p_k - \theta_i f(Q_k, \bar{C_k}) \tag{2.3}$$

Therefore, if $p_m - p_k \leq \theta_i \left(f(Q_k, \bar{C_k}) - f(Q_m, \bar{C_m})\right)$ and for monotonic $f(Q_m, \bar{C_m})$, class m will be preferred over k if

$$\begin{aligned} &\theta_i \geq (p_m - p_k)/(f(Q_m, \bar{C_m}) - f(Q_k, \bar{C_k})), \text{when } k > m \\ &\text{and if} \\ &\theta_i \leq (p_m - p_k)/(f(Q_m, \bar{C_m}) - f(Q_k, \bar{C_k})), \text{when } k < m \end{aligned} \tag{2.4}$$

This creates thresholds of θ values, $\theta_1 > \theta_2 > ... > \theta_M > \theta_{M+1} = 0$, such that at equilibrium, for $1 \leq m \leq M$, $\forall \theta_i \in (\theta_{m+1}, \theta_m)$, class m is chosen, while no class is preferred for $\theta_i > \theta_1$. The thresholds $\theta_1, ..., \theta_{M+1}$ are defined by using the fact that, at any of these specific threshold, a UE is indifferent between choosing one of the two adjacent classes. A UE, at threshold θ_1, is also indifferent between using the provided service or not.

We let the congestion perception function $f(Q_m, \bar{C_m})$ of the UEs to be:

$$f(Q_m, \bar{C_m}) = \frac{Q_m}{\bar{C_m}/C_m} \tag{2.5}$$

We introduce a dynamic pricing scheme for each access class m. The maximum price for each class $p_m^{\max}$, is set for accessing class m when its total capacity C_m is allocated, and the minimum price $p_m^{\min}$ is set when the total capacity of class m is available. The price as a function of available capacity $\bar{C_m}$ is expressed in (4.4).

$$p_m = \max\left(p_m^{\min}, p_m^{\max}\left(1 - \frac{\bar{C_m}}{C_m}\right)\right) \tag{2.6}$$

Regarding the mobility model of the UEs, we consider that they follow routes of diverse connectivity conditions to the available radio access technologies, according to the model proposed in [81]. In Figure 2.2, we present the states of the Markov model for the mobility of a UE. A UE in State 0 will pass through the multi-RAT area (State 1) that we focus on with probability p_{01}, it will stay in State 1 with probability p_{11} and will leave the multi-RAT area with probability p_{12}. A UE starting from State 0 may not pass through the multi-RAT environment with probability p_{03} and stay at State 0 with probability p_{00}. Following, we provide the system architecture that was considered for the evaluation of our dynamic PMP scheme, and the algorithms designed for the multi-RAT operation.

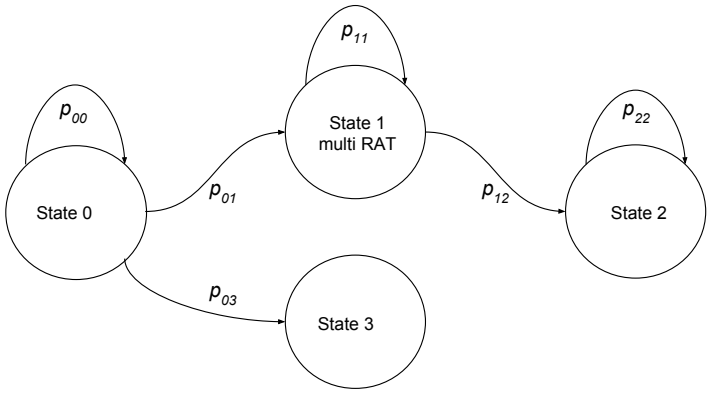


FIGURE 2.2: Markov Model States for the mobility of the UEs.

2.4 System Architecture

In this section, we describe the system architecture and the components that are taken into consideration from the proposed solution. We present the configurations and parameters for both simulations and real-world experiments. Then, we describe the procedures of the system model that take place both at the Core Netowrk and the UEs of the Multi-RAT network.

2.4.1 Experimental setup

The proposed scheme is considered for three available radio access technologies, namely 3G, 4G and WiFi, and is evaluated by means of extensive simulations. Taking into consideration the significant shortfalls of simulations, identified in [63], in comparison to real-life experimentation, we provide results based also on testbed experimentation implemented on the Future Internet (FI) facility provided by NITOS [83]. NITOS is a heterogeneous testbed, located in the premises of University of Thessaly, in Volos, Greece. The rich heterogeneity of resources allows us to conduct the designated experiments. We employ an LTE base station of NITOS, along with a UMTS femto-cell and the respective Core Network [72]. We use a testbed node as a WiFi Access Point, located inside the coverage of both LTE and UMTS, and six nodes as the UEs located inside the Multi-RAT system. The overall topology that we use for our experiments is depicted in Figure 2.3.

For the simulations we consider a basic system setup and demonstrate how the proposed pricing policy performs in certain representative scenarios. The mobile service provider has 3 classes of RAT and specifically cellular 3G, 4G and WiFi services. We assume that the maximum speed for each RAT is 42.4 Mbps (UMTS/HSPA+), 150 Mbps (IEEE802.11n SISO) and 300 Mbps (LTE Cat4 2x2 MIMO).

2.4.2 Algorithms Design

In order to evaluate the proposed pricing scheme, we came up with two algorithms, running at the UEs and the Core Network. In **Algorithm 1**, we present the operation of a UE that enters or is already situated in the multi-RAT environment. Initially, a UE receives a system report message containing the price, the available capacity and the mass of connected users to every RAT, and it calculates its utility for each class. If it decides to change or connect to a RAT, it sends a change/connect message to

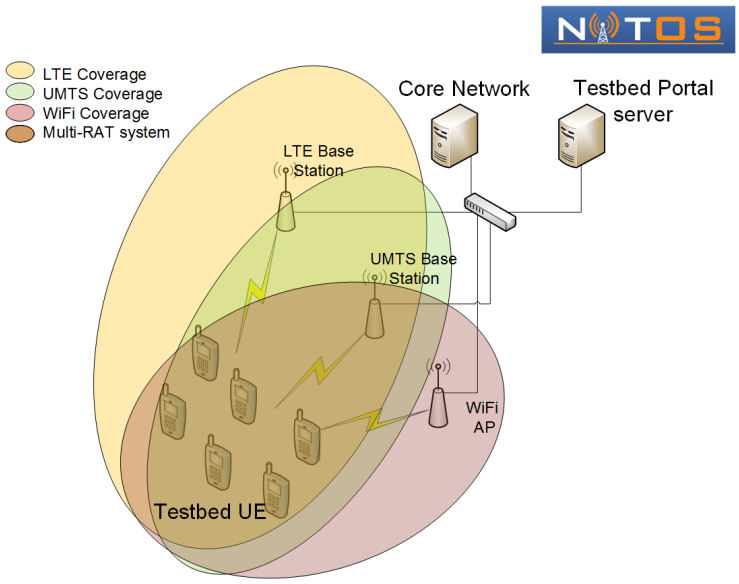


FIGURE 2.3: NITOS Testbed topology used for the experiments

Algorithm 1 Algorithm for UEs entering or already connected to the multi-RAT environment

```
Receive system report (p_m, C̄_m, Q_m) ∀ RAT
Calculate Utility Function for each RAT
if UE decides to change/connect to RAT then
    Send change/connect message to Core Network
else
    Send OK to Core Network
end if
while 1 do
    if UE decides to leave State 1 then
        Send leave message to Core Network
    else
        Wait for system report
        Calculate Utility Function for each RAT
        if UE decides to change RAT then
            Send change message to Core Network
        else
            Send OK to Core Network
        end if
    end if
end while
```

the Core Network. When a UE is connected and until it leaves State 1, if it receives a new system report, it recalculates its utility function for each class and examines if a change of RAT maximizes its utility. Based on its decision, the UE sends a message to the Core network to inform for a change (or for no change) and continues its operation as long as it stays in State 1.

In **Algorithm 2**, we provide the functionality of the Core Network. At first, the Core Network randomly distributes the UEs situated in State 1 to the available RATs, and it calculates the prices, the available capacities and the mass of connected users

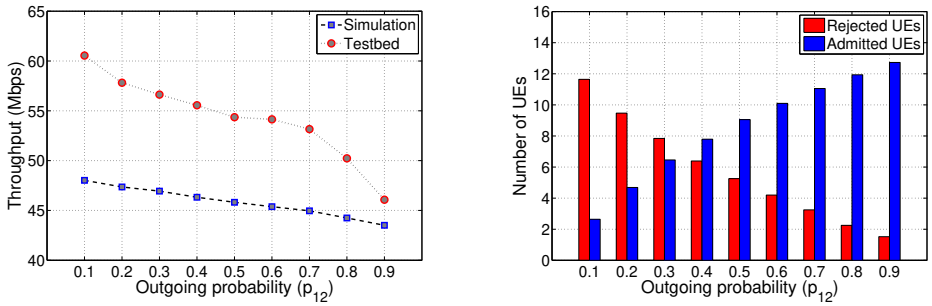
Algorithm 2 Algorithm running on the Core Network

```
Assign randomly UEs to 3 RATs
Calculate prices, available capacities and mass of users
Create system report (p_m, C̄_m, Q_m) ∀ RAT
for UEs in system do
    Send system report to UE
    Wait for response
    Update system report
end for
if System is stable then
    Continue
else
    Go to 4
end if
while 1 do
    Wait update from UEs in State 1
    Wait for new UEs
    if Available capacity then
        Assign new UE to RAT
    else
        Deny access to new UE
    end if
    Calculate prices, available capacities and mass of users
    Create system report (p_m, C̄_m, Q_m) ∀ RAT
    for UEs in system do
        Send system report to UE
        Wait for response
        Update system report
    end for
    if System is stable then
        Continue
    else
        Go to 24
    end if
end while
```

to every RAT. Thereafter, it sends an updated system report to each UE sequentially, and following it examines if the system is stable. The system is considered stable, when no UE desires to switch to another RAT. In the case when a UE leaves the multi-RAT, the Core Network updates the system values and communicates them to the State 1 UEs for further calculation of their utility functions, in order to determine whether they will change their RAT. In the case that a UE enters State 1, the Core Network calculates the available capacity to decide if it can serve its needs, otherwise the access is denied to the specific UE.

2.5 System Evaluation

For the evaluation of our proposed policy, we employ both simulation methods, as well as real testbed experiments. For our simulations, we evaluate our model for 10 users, who are initially placed in State 1 and 20 users in State 0. In order to validate



(A) Throughput performance for UEs exiting the system

(B) Number of UEs admitted and denied by the system

FIGURE 2.4: Experiment results for UEs exiting the multi-RAT system

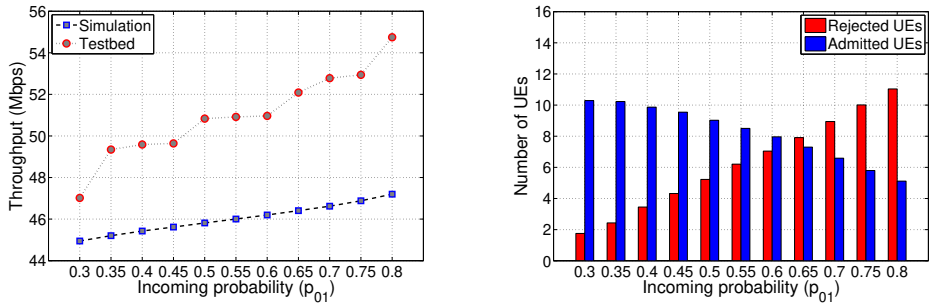
our simulation results, we cross-reference them with results that we received from executing our proposed policy under a real testbed environment, with initially 2 users in State 1 and 4 users in State 0.

For the testbed tests, we logged RSSI and RSRP values for the LTE network equal to -53 dBm and -76 dBm respectively. Similar values where observed for the UMTS and WiFi networks as well, inside the multi-RAT environment. Rate adaptation algorithms were disabled, and set to the highest Modulation and Coding Scheme for all the technologies involved, meeting the same maximum throughput requirements as set in our simulator. Moreover, we conduct our experiments in an isolated environment of external interference. We measure the capacity of our links by using packet sizes of 1500 bytes. The maximum and minimum prices of each class are selected to be proportional to the maximum available capacity of each RAT. We use a static probability p_{03} equal to 0.2, representing the probability that UEs on State 0 follow a route outside our system across all our experiments (simulations and testbed experiments). In other words, the selected mobility pattern includes a more common passage from the multi-RAT State 1 than avoiding it through State 3.

We organize our evaluation in three different experiments, during which the prices are constantly recalculated based on our model, and transmitted to the UEs of the network. The first experiment targets to monitor the throughput and acceptance ratio of the UEs in the multi-RAT system, when tuning the probability of exiting the system. The second experiment evaluates the same metrics when configuring the probability of a UE entering the multi-RAT system. Finally, the third experiment details our results for different congestion sensitivities of the UEs.

2.5.1 Experiments for UEs exiting the system

In this experiment, we seek to investigate the impact of different exiting mobility patterns, represented by the probability p_{12}, on the average throughput of the system. Furthermore, we examine how tuning this probability affects the number of incoming users from State 0, as well the number of users whose the access to State 1 was denied from the Core Network due to insufficient available capacity. For this experiment we configure the probabilities for the State 0 to have the following values: $p_{00} = 0.3$ and $p_{01} = 0.5$. In addition, we setup the highest capacity demand of each UE equal to 1/4 of the RAT with the highest capacity, in our case the LTE network. Figure 2.4a demonstrates the average achieved throughput for various values of the exiting probability p_{12} used to express whether a UE leaves the multi-RAT system or not. In our simulations, we observe that as the probability of leaving the system is



(A) Throughput performance for UEs entering the system

(B) Number of UEs admitted and denied by the system

FIGURE 2.5: Experiment results for UEs entering the multi-RAT system

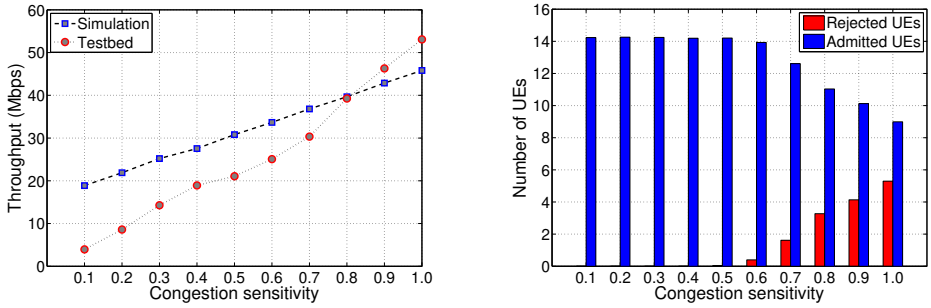
increasing, the total throughput is decreasing. This happens due to the fact that UEs stay at State 1 for a shorter period. On the other hand, the number of UEs that are not accepted to State 1 is decreasing as the probability p_{12} increases, meaning that more UEs enter the system as shown in Figure 2.4b.

Similarly, for our testbed experiments, we observe a similar decreasing pattern as the outgoing probability rises. We would expect that the testbed experimentation would generate results with lower average throughput compared to simulations. Nevertheless, as the experiments are run using a lower number of UEs, we receive marginally higher average throughput.

2.5.2 Experiments for UEs entering the system

In this scenario, extensive throughput experiments were conducted under varying values of the probability p_{01}. The total throughput performance is investigated for different mobility patterns of UEs entering the multi-RAT system. The probabilities in this scenario regarding State 1 are the following: $p_{11} = p_{12} = 0.5$, meaning that each existing UE inside the multi-RAT has a probability of 0.5 to either stay in or exit the system. The highest capacity demand of each UE stays the same as in the previous experiment. We notice that while p_{01} increases, the same occurs to the average system throughput as illustrated in Figure 2.5a. Moreover, increasing the entering probability causes more UEs to be served by the multi-RAT system as observed in Figure 2.5b. In parallel, the total number of UEs choosing a transition to State 1 rises (comprising of both admitted and denied UEs). Of course, the more UEs select to enter the multi-RAT system, the more are denied access due to lack of available capacity.

Regarding the testbed evaluation, we use a similar setup as for the aforementioned experiments. We observe that as the incoming probability rises, so does the average throughput that the UEs in the multi-RAT system achieve. Higher average throughput denoted in the testbed experiments, in comparison to simulation results, is mirroring the lower number of UEs used for our experiments.



(A) Throughput performance with respect to different congestion sensitivities

(B) UEs admitted and denied with respect to different congestion sensitivities

FIGURE 2.6: Experiment results subject to different congestion sensitivities for the UEs' application

2.5.3 Experiments for different congestion sensitivities

Through this set of experiments, we aim at investigating the impact of different QoS applications to our proposed PMP scheme. More specifically, we characterize different classes of applications by their congestion sensitivity, running inside the multi-RAT environment we focus on. In order to examine how applications with different demands affect the average throughput of the system, we first configure the probabilities of State 0 and 1 with the following values: $p_{00} = 0.3$ and $p_{01} = 0.5$ and $p_{11} = p_{12} = 0.5$. This means that each new UE stays in State 0 with probability 0.3, enters the multi-RAT system with probability 0.5, and follows a route outside our system with probability 0.2. Each existing UE inside State 1 has a probability of 0.5 to either remain in the system or exit it. By tuning different congestion sensitivities, we simulate applications with different QoS demands in the system. A UE running an application with the maximum congestion sensitivity is depicted in Figure 2.6a with the value 1.0. In our results, the throughput received from such an application with high congestion sensitivity is equal to the 1/4 of the total LTE's bandwidth. We have also assumed that the values of the congestion sensitivity of the UEs in the multi-RAT are uniformly distributed. Figure 2.6a depicts the average throughput achieved for different values of congestion sensitivity. We observe that for low bandwidth demands, all new UEs are accepted and served by the system but the average throughput performance is low, ought to the low requested demands. As the congestion sensitivity increases, the average throughput also increases, which means that the available capacity is decreasing. Consequently, the system is starting to deny access to new UEs, as it is shown Figure 2.6b. Similarly for the testbed experiments, we observe a pattern that follows the simulation trends for the overall average throughput performance.

2.6 Chapter Conclusion

In this chapter, we study the resource allocation problem in a multi-RAT environment. We proposed the utilization of an extended Paris Metro Pricing scheme with dynamic prices, as a policy for selecting a RAT when operating inside a Heterogeneous network. We formulated the problem and defined the utility functions of each client for accessing the target RAT. For the mobility of the clients, a Markov model was utilized, where the probabilities for changing states were configurable thus enabling us to investigate several mobility scenarios. For the evaluation of our

scheme, simulations were employed in order to validate it and then we utilized the NITOS testbed so as to obtain results from a real-world environment. Three different set of experiments were conducted towards the evaluation of our dynamic PMP scheme, providing results on the average throughput, the acceptance capability of the incoming users and how these performance metrics are affected under different mobility conditions. In the next chapter, we are extending the current resource allocation scheme, by enabling a UE to be concurrently connected to multiple RATs, and to interchangeably select its access technology based on the current pricing policies of the system.

Chapter 3

Resource Allocation in 5G HetNets; Scaling for Multi-homed Clients and Multiple Traffic Classes

Contents

3.1 Chapter Introduction

Ultra Dense Heterogeneous Networks (HetNets) are expected to bring numerous advantages for the mobile network operators (MNOs) of the next generation cellular networks, by offering enhanced network capacity and diverse technologies for serving the end-users. Through enhanced spectral techniques and spectrum coordination among the available cells, 5G networks will utilize bands ranging from the sub6-GHz to cm- and mm-Wave bands. Yet, this type of equipment upgrade for the Radio Access Network (RAN) incurs extra deployment and operation costs for the MNOs. When considering that the Average Revenue Per User (ARPU) is either flat or even slowly decreasing [76], it becomes clear that cost efficient techniques are required in order to keep in pace with the demand-driven evolution in the cellular technology.

Nevertheless, through the utilization of off-the-shelf components operating in multiple bands and with different technologies, HetNets can be formed with low CAPEX and OPEX costs. Forming HetNets for adding to the overall network capacity has existed for legacy technologies as well, (e.g. UMTS, LTE) through the addition of low energy consuming devices with lower coverage in the network (e.g. femto-/pico-cells). Recent efforts have also focused on the inter-networking of cellular technologies with WiFi as well, as a means to add-up to the offered network capacity with low-cost solutions (e.g. through the LTE-WiFi Aggregation Adaptation

Protocol [4]). The incorporation of the new spectrum that 5G is expected to operate, creates a complex ecosystem for allocating the operator served clients to each of the offered technologies, while meeting diverse demands for network capacity.

This technology allocation problem becomes even more complicated when considering that the perceived Quality of Experience (QoE) highly depends on the application that is running on the network UE. Different applications have versatile demands for the network capacity; UHD video requires more than 25Mbps to be dedicated for the application, whereas an email application is only served as background traffic. Based on the existing technology offering for running multi-homed end-clients (e.g. MPTCP, using SDN and software switches [93]), a network client might benefit from having simultaneously several connections of different technologies active to the Internet. In such an environment, efficient mapping of each application to the available technologies might significantly enhance the perceived QoE, whereas alleviate the burden placed on the operator serving multiple data-intensive traffic streams. Nevertheless, multiple traffic classes for the applications add to the complexity of the network resources allocation to the served clients.

The main questions that arise in such heterogeneous environments with multiple RATs being offered to the network UEs, and multiple traffic classes per UE are:

1. Which technology should a UE select for the different types of traffic.
2. When should a UE switch the serving network for a specific traffic class.
3. How should an operator charge each RAT.
4. How these decisions affect the overall system stability.

In order to answer these questions, we extend our work presented in Chapter 2, where we introduced a network selection scheme based on the Paris Metro Pricing (PMP), a service differentiation scheme that was first used in Paris metro to give to its passengers the ability to opt for less congested wagons. In this chapter, we present a system model that takes into consideration multiple traffic classes concurrently utilized by each UE, and multiple networks being offered by the MNO, forming a Multi-RAT system. We evaluate the performance and stability of the system through extensive experimentation over multiple technologies in an open wireless testbed.

The rest of this chapter is organized as follows: Section 3.2 presents relevant works on the RAT selection policies and methodologies. In Section 3.3 we introduce our system model. Section 3.4 includes the description of our system architecture, and in Section 3.5 we showcase our experimental findings. Finally, in Section 3.6 we conclude our work.

3.2 Related Work

Different access schemes have been proposed in literature for exploiting the coexistence of heterogeneous wireless technologies for improving the Quality of Service (QoS) and Experience (QoE) of the mobile users. In [112], a survey on the different models for network selection in HetNets is provided, with the solutions being classified based on the proposed utility functions and system models. Similarly, in [80], authors classify the respective algorithms based on the location where the decision making network selection components are running, as either partially or fully distributed. In [20], a fully distributed access point selection algorithm is presented based on no-regret learning. Through the application of this scheme, the system

is able to reach a correlated equilibrium state. Authors in [84], formulate the RAT selection problem using a dynamic evolutionary game and introduce a centralized algorithm based on reinforcement learning. In [8], the RAT selection problem is modeled as a non-cooperative game, and is evaluated for its convergence, efficiency, and practicality. Through their approach, each user tries to selfishly maximize its own throughput, while the impact of a user's decisions on other users performance and the convergence of the system to Nash equilibria is investigated. The authors conclude that an improvement path can be repeated infinitely with a mixture of classes. As network convergence is the target of these algorithms, authors in [77] discuss the convergence properties of network selection games. The network selection process is studied as a non-cooperative game, and is evaluated for the cases where each client uses its own preference to select a network, and for a combination of client and network preferences to arrive at pairings.

Contrary to the majority of works that deal with the maximization of the user throughput, authors in [31] propose the user demand-centric optimization, where users seek to maximize quality of experience (QoE). Their research validates the existence of user demand diversity gain and the effectiveness of their learning algorithm in improving the system efficiency and QoE fairness. In [11], the authors utilize a reliable low-latency Fiber Backhaul sharing and WiFi Offloading and evaluate the maximum aggregate throughput and the delay performance of FiWi enhanced LTE-A HetNets. Authors in [**Naghavi2016**] consider a heterogeneous cellular network where each user chooses among multiple access technologies. The competition of the users is modeled as an incomplete information game where players are not aware of other players' actions. An incentive mechanism that aims to motivate WiFi Access Points (APs) to participate in heterogeneous networks, by providing an access class to the existing cellular infrastructure, is proposed in [58]. The pricing strategy for the inclusion of third party WiFi APs is formulated as a Stackelberg game between the mobile network provider and the third party WiFi APs. The authors in [86], investigate the impact of a cooperative switching off in multi-operator shared HetNets in terms of cost allocation in the MNOs.

The authors in [18] consider a general model of congestion externality for the PMP and investigate the conditions of congestion functions that guarantee the viability of the PMP scheme. Similarly, in [89] we propose and evaluate a dynamic pricing algorithm for HetNets, based on the PMP scheme. In this chapter, we build on our prior experience and further extend it to MATCH (**M**ultiple **A**ccess for multiple **T**raffic **C**lasses in 5G **H**etNets), aiming to include multiple traffic classes per each UE, corresponding to the diverse network communication needs that applications serving the end-user are requiring. We further extend the system model with multi-homed UEs and with the inclusion of higher throughput 5G technologies (e.g. WiGig) and evaluate the proposed pricing scheme with a real testbed evaluation.

3.3 MATCH System Model

We consider a heterogeneous environment with M classes of available radio access technologies that belong to the same cellular service provider. Each class m refers to a different radio access technology (e.g. 4G, 3G, WiFi, mmWave) with capacity C_m, resulting to a total system capacity equal to $C = \sum_{m=1}^{M} C_m$. We assume that there are L types of user traffic classes and each UE_i may have traffic demands for these types of traffic. The traffic of each UE_i is characterized by the vector $\boldsymbol{\theta}_i = \{\theta_i^1, ..., \theta_i^L\}$, where $\theta_i^l = r_i^l / w_i$ represents the sensitivity of UE_i's traffic of type l, with r_i^l being the

data rate demand of UE_i for its traffic class l, and w_i the normalized spectrum efficiency of UE_i, with $w_i \in (0,1]$ for $i = (1,...,N)$, which is used to abstract the physical layer for the channels of the available RATs, including the frequency selectivity due to transmissions in different frequency bands. As we focus on access layer decisions, we provide w_i as a plug-in parameter to our dynamic pricing scheme, available for physical layer analysis. θ_i^l represents the ability of UE_i's traffic of type l to adapt easily to changes in the network conditions, while still meeting specific QoS requirements and maintaining the QoE for the user. The mass of traffic classes allocated to RAT class m is denoted by Q_m, and without loss of generality we assume that the total mass of traffic classes of the UEs is equal to 1. The mass of traffic classes that are not allocated to a RAT class is equal to $Q_0 = 1 - \sum_{m=1}^{M} Q_m$.

We focus on UEs that are under the coverage of all provided radio access technologies as depicted in Figure 3.2. Each traffic class l of UE_i is allocated to a RAT class m such that the corresponding element $b_i^{m,l}$ of an $M \times L$ matrix B_i is equal to θ_i^l, if its traffic class l is allocated to the RAT class m, and 0 otherwise. Thus, the traffic allocation matrix B_i of UE_i, with columns representing the UE's traffic classes and lines the available RAT classes, is expressed as:

$$B_i = \begin{pmatrix} \theta_i^1 & 0 & 0 & \cdots & 0 \\ 0 & \theta_i^2 & \theta_i^3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \theta_i^L \end{pmatrix} \tag{3.1}$$

We define as $\mathbf{b}_i$, a vector of size $1 \times M$, where each element b_i^m is the sum of each row of B_i. Thus, $b_i^m = \sum_{l=1}^{L} b_i^{m,l}$. We also define $\mathbf{a}_i$, a RAT access index vector of size $1 \times M$, where $a_i^m = 1$ if $b_i^m > 0$ and $a_i^m = 0$ if $b_i^m = 0$.

A UE_i with congestion sensitivity $\boldsymbol{\theta}_i$, using the access classes indicated by $\mathbf{a}_i$ has a utility equal to:

$$U_i(B_i) = V(\mathbf{1} \cdot \mathbf{a}_i) - \mathbf{a}_i \cdot \mathbf{p} - \boldsymbol{\theta}_i \cdot \mathbf{f} \tag{3.2}$$

where V is a flat-rate valuation of accessing a class of the multi-RAT service and $\mathbf{1} \cdot \mathbf{a}_i$ is the number of concurrently accessed RAT classes by UE_i. The access charge vector $\mathbf{p} = \{p_1, ..., p_M\}$ represents the price of each RAT at the time a UE is deciding to which RAT classes it will allocate its traffic classes. The vector of functions $\mathbf{f} = \{f_1(Q_1, \bar{C}_1), ..., f_M(Q_M, \bar{C}_M)\}$, represents the perceived congestion of the M available RAT classes by all UEs in the HetNet. Thus, $f_m(Q_m, \bar{C}_m)$ is a function for the perceived congestion at class m, and the available capacity of access class m is denoted by $\bar{C}_m$. We assume that $V \geq p_1 > p_2 > ... > p_M$ and therefore a UE_i with traffic classes of sensitivity $\boldsymbol{\theta}_i$ will choose, in a decomposed approach for each traffic class l, the service class that maximizes its discrete traffic class utility:

$$U_i^l(m) = V - p_m - \theta_i^l f_m(Q_m, \bar{C}_m) \tag{3.3}$$

such that:

$$m(\theta_i) = \underset{1 \leq m \leq M}{\operatorname{argmax}} \, U_i^l(m) \tag{3.4}$$

or no service if $U_i^l(m) = 0, \forall\ m \in (1,...,M)$ and $\forall\ l \in (1,...,L)$. This leads to a two-level (Stackelberg) game, where the provider first decides the prices per access class, as a function of the already allocated capacity per access technology, and then the UEs distribute their traffic classes over RAT classes, selecting the most appropriate for each traffic class. The provider can play by anticipating the distribution of the UEs traffic classes over the provided service classes. The distribution of traffic

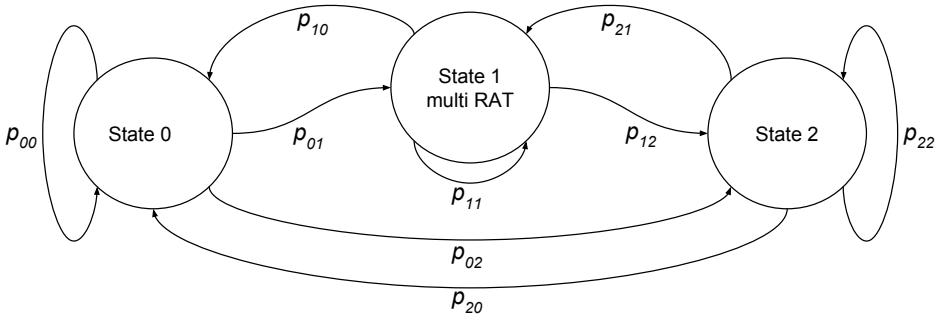


FIGURE 3.1: Mobility States of the UEs.

classes over the provided access technologies will be a Wardrop equilibrium [113], at which no UE will have an interest in changing its traffic classes allocation to available access classes. When equilibrium of distribution of traffic classes over available access classes is reached, a given traffic class l of UE_i of sensitivity θ_i^l will prefer class m over k if

$$V - p_m - \theta_i^l f_m(Q_m, \bar{C}_m) \geq V - p_k - \theta_i^l f_k(Q_k, \bar{C}_k) \tag{3.5}$$

Therefore, if $p_m - p_k \leq \theta_i^l \left(f_k(Q_k, \bar{C}_k) - f_m(Q_m, \bar{C}_m)\right)$ and for monotonic congestion perception functions $f_m(Q_m, \bar{C}_m)$, class m will be preferred over k if

$$\theta_i^l \geq (p_m - p_k)/(f_m(Q_m, \bar{C}_m) - f_k(Q_k, \bar{C}_k)), \text{when } k > m$$

and if

$$\theta_i^l \leq (p_m - p_k)/(f_m(Q_m, \bar{C}_m) - f_k(Q_k, \bar{C}_k)), \text{when } k < m$$

This creates thresholds of θ values, $\theta_1 > \theta_2 > ... > \theta_M > \theta_{M+1} = 0$, such that at equilibrium, for $1 \leq m \leq M$, $\forall\ \theta_i^l \in (\theta_{m+1}, \theta_m)$, class m is chosen, while no class is preferred for $\theta_i^l > \theta_1$. The thresholds $\theta_1, ..., \theta_{M+1}$ are defined by using the fact that, at any of these specific threshold, a UE is indifferent between choosing one of the two adjacent access classes for a specific traffic class. A traffic class, at threshold θ_1, is also indifferent between using the provided service or not.

We let the congestion perception functions $f_m(Q_m, \bar{C}_m), \forall\ m \in (1, ..., M)$ to be:

$$f_m(Q_m, \bar{C}_m) = \frac{Q_m}{\bar{C}_m / C_m} \tag{3.6}$$

We introduce a dynamic pricing scheme for each access class m. The maximum price for each class $p_m^{\max}$, is set for accessing class m when its total capacity C_m is allocated, and the minimum price $p_m^{\min}$ is set when the total capacity of class m is available. The price as a function of available capacity $\bar{C}_m$ is expressed in (4.4).

$$p_m = \max\left(p_m^{\min}, p_m^{\max}\left(1 - \frac{\bar{C}_m}{C_m}\right)\right) \tag{3.7}$$

Regarding the mobility model of the UEs, we consider that they follow routes of diverse connectivity conditions to the available radio access technologies, inspired by the model proposed in [81]. In Figure 3.1, we present the states of the Markov model for the mobility of a UE. A UE in State 0 will pass through the multi-RAT area (State 1) that we focus on with probability p_{01}, will stay in State 1 with probability

p_{11} and will leave the multi-RAT area with probability p_{12}. A UE starting from State 0 may not pass through the multi-RAT environment with probability p_{02} and stays at State 0 with probability p_{00}. We consider the return probabilities from State 2 to State 1 and to State 0, equal to p_{21} and p_{10} respectively.

In the following sections, we provide an approach to evaluate the distributed solution we provide with MATCH. Based on the system model, we port it to run in a distributed manner on all the network, i.e. the UEs and the Core Network. Using the outcomes of the utility functions running on each network UE, a Core Network controller is able to suggest the technologies and list the prices for using them.

3.4 System Architecture

In this section, we detail the components and the architecture of our testbed setup used for the evaluation of our proposed framework. Afterwards, the algorithms of the system model that are running at the Core Network and the UE side are presented.

3.4.1 Experimental Setup

For the evaluation of the proposed scheme, we use four RATs, each one with different characteristics (3G, WiFi, 4G and mmWave), and four different traffic classes available at each network UE. Each Traffic Class (TC) is categorized by its data rate demands, ranging from Traffic Class 0 $(TC0)$ designating applications that require Best Effort connectivity (e.g. email) to Traffic Class 3 $(TC3)$ designating time critical/bandwidth intensive applications (e.g. UHD video stream).

We provide experimental results based on the evaluation of the proposed scheme over the NITOS Future Internet (FI) facility [83]. NITOS is a heterogeneous testbed, located in the premises of University of Thessaly, in Volos, Greece. The rich heterogeneity of its provided resources allows us to conduct the designated experiments. We employ an LTE base station, along with a UMTS femto-cell and the respective Core Network. We use a testbed node as a WiFi Access Point, located inside the coverage of both LTE and UMTS. NITOS has been upgraded recently with six mmWave WiGig radio units [13], that are reachable by all the testbed nodes. In order to include WiGig in the Multi-RAT technologies, we use a pair of WiGig nodes, that are reachable and interchangeably usable by the UEs involved in the experiment. The overall topology that we use for our experiments is depicted in Figure 3.2.

3.4.2 Algorithms Design

In order to port the system model setup over the testbed equipment, we came up with two algorithms for the UEs and the Core Network. Algorithm 3 is running on the distributed UEs of the system. When a UE enters the multi-RAT environment, it receives a system report message containing the available capacity, the current price, and the load of connected users for every RAT in the system. Following, the UE calculates its utility for its Traffic Classes, based on all the available RATs. Subsequently, the Core Network controller is informed on the UE's decisions on which RAT it decides to allocate its Traffic Classes (TCs). During the time that the UE spends in the multi-RAT system, it is waiting to receive a new system report issued by the Core Network periodically. In such a case, and as long as the UE stays in State 1, it recalculates the utilities and selects the RAT for each TC.

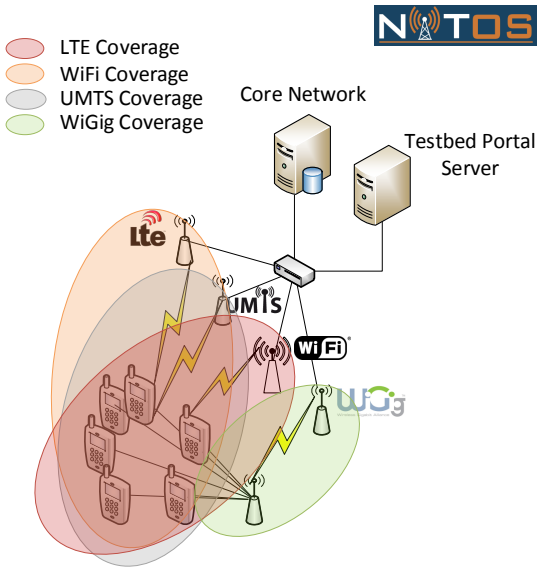


FIGURE 3.2: Experimental topology for the evaluation of the scheme in the NITOS testbed

Algorithm 3 Algorithm running on the UE side

```
if UE is connected to multiRAT then
    Receive system report (p_m, C̄_m, Q_m) ∀ RAT
    Calculate the Utility for each TC per RAT
    UE sends change/OK message to Core Network
else
    if UE decides to connect to multiRAT then
        Send connect message to Core Network
    end if
end if
while 1 do
    if UE decides to leave State 1 then
        Send leave message to Core Network
    else
        Wait for system report
        Calculate Utility Function for each TC per RAT
        UE sends change/OK message to Core Network
    end if
end while
```

The design of the algorithm running at the Core Network is presented in Algorithm 4. The Core Network controller initializes the multi-RAT system by randomly distributing the State 1 UEs' TCs to the available RATs. Thereafter, it calculates the available capacity, price, and the mass of TCs per RAT. Based on the applied ordering of users, the controller starts sending a system report to each UE sequentially and waits for their response. Based on their response, it determines if the changes of their TCs are applicable or not and proceeds accordingly. In the case that a UE enters State 1, the controller determines if the UE can be served by the system and approves or denies access. For every change in the multi-RAT system (UEs entering/leaving State 1), the controller updates the system values and communicates them to the

Algorithm 4 Algorithm running on the Core Network

```
Assign randomly the UEs' TCs to 4 RATs
Calculate prices, available capacities and mass of TCs
Create system report (p_m, C̄_m, Q_m) ∀ RAT
Create an order of UEs based on (Data Rate Demand/Spectrum Efficiency/Sensitivity)
for UEs in system do
    Send system report to UE
    Wait for UE response
    Update system report
end for
while 1 do
    Wait update from UEs in State 1
    Wait for new UEs
    if Available capacity then
        Assign new UE's TCs to RATs
    else
        Deny access to new UE
    end if
    Calculate prices, available capacities and mass of TCs
    Create system report (p_m, C̄_m, Q_m) ∀ RAT
    Create an order of UEs based on
(Data Rate Demand/Spectrum Efficiency/Sensitivity)
    for UEs in system do
        Send system report to UE
        Wait for response
        Update system report
    end for
end while
```

State 1 UEs for further calculation of their utility functions (as shown in Algorithm 3). As seen in Algorithm 4, the Core Network policy for determining the system report depends on the ordering of the UEs based on their data rate demands, spectrum efficiency, and sensitivity. Hence, in the following section we provide results on three different UE ordering policies.

3.5 System Evaluation

In this section we provide the experimental evaluation of our proposed policy. As the testbed is organized in an RF-isolated setup, we are able to reproduce each experiment. We showcase our experimental results with a resolution of 10 times per experiment, with 6 UEs available in the Multi-RAT system.

Based on the insights from our previous work [89], we evaluate our proposed scheme for three different polling policies.

1. **Policy 1**: The controller is ordering the State 1 UEs based on their data rate demands for all TCs ($\sum_{l=1}^{4} r_i^l$) in descending order.
2. **Policy 2**: The ordering is based on the normalized spectrum efficiency of the UEs (w_i).
3. **Policy 3**: The ordering is based on the UEs' sensitivity to the changes of the network conditions ($\boldsymbol{\theta}_i$).

In order for MATCH to be evaluated in a real world environment, we determined a setup of clients in the testbed complying with the needs of our model. The rate adaptation algorithms were disabled for all the technologies and the highest Modulation and Coding Scheme was configured for all of them. We measure the capacity of our links by using frames of 1500 bytes and we set the maximum achieved speed for the used technologies to be 42 Mbps for UMTS/HSPA+, 70 Mbps for LTE 10MHz 2x2 MIMO, 130 Mbps for IEEE802.11n 20MHz 2x2 MIMO and 1 Gbps for WiGig. For the configuration of the TCs of each UE, we assume 4 different classes with limits in the lowest and highest throughput that the UEs request from the network. Each UE selects uniformly the throughput value from this range, configured as follows: 1) Background class (0.3-1Mbps), 2) Interactive class (1.1-2.5 Mbps), 3) Streaming class (2.6-8Mbps) and 4) Conversational class (8.1-25Mbps). The limits have been set based on [105].

We configure the probabilities of State 0 with p_{02} to be equal to 0.2, representing the probability that UEs on State 0 follow a route outside our system across all our experiments. The probability of a UE in State 0 to enter the Multi-RAT system (p_{01}) is equal to 0.5 and thus the probability of staying in the same state (p_{00}) is 0.3. For State 1, the probability a UE exits the system going to State 2 (p_{12}) is configured to 0.45 and to remain in the same state (p_{11}) to 0.5, meaning that the probability of returning to State 0 (p_{10}) is 0.05. For State 2, we set the probability of a UE to remain in this state (p_{22}) equal to 0.9 and the probabilities p_{20} and p_{21} equal to 0.05.

The scenario of the experiments that we present in this section is the following: The experiment starts with UE0 and UE1 in State 1, while UE2, UE3, and UE4 will be entering the system later. UE5 chose the direct route to State 2 and remained there.

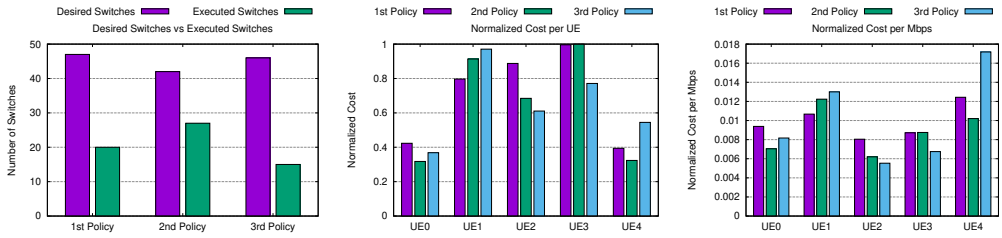
3.5.1 Framework benchmarking

For the experiment under consideration the aggregate achieved throughput for all traffic classes for each UE in the system is presented in Table 4.5. We ran the experiment for each of the three cases ten times and we present the averaging results. Figure 3.3 is showcasing our averaged results from running the aforementioned use case in the testbed. Figure 3.3a demonstrates how many times the UEs determined a RAT change for their TCs during the experiment (Desired Switches) versus how many times they accomplished to do so (Executed Switches). As Algorithm 4 recalculates capacities, prices and mass of TCs per each RAT, it may infer that the Utility calculated per each UE for a RAT change may not be valid, and subsequently denies the RAT selection. As it is shown, when the criterion for the ordering is the normalized spectrum efficiency (2nd Policy), the UEs have more freedom to change RATs compared to 1st and 3rd case. When applying the 3rd Policy, the UE ordering renders any RAT selections rather difficult.

TABLE 3.1: Total Achieved Throughput per UE

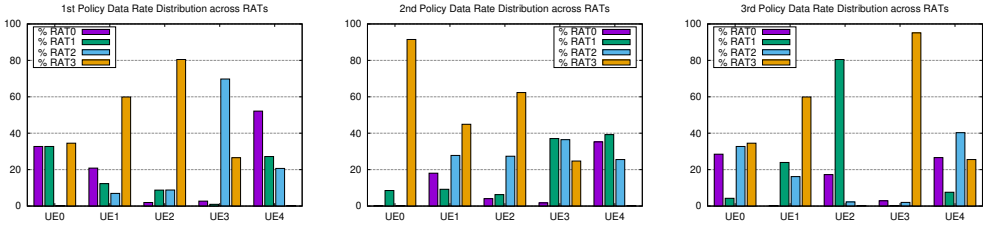
UE ID	Throughput (Mbps)
UE0	45.09
UE1	74.65
UE2	110.36
UE3	114.27
UE4	31.72

Following, we examine how the different policies affect the charges of the multi-RAT system. In order to compare the three policies, we normalize the costs as shown



(A) Switches performed per each policy (B) Normalized cost for the three policies (C) Normalized cost per Mbps

FIGURE 3.3: Experimental evaluation of the proposed resource allocation scheme



(A) 1st Policy: Data Rate ordering (B) 2nd Policy: Spectrum ordering (C) 3rd Policy: Sensitivity ordering

FIGURE 3.4: Data Rate distribution of each client across the System's RATs for the different ordering policies

in Figure 3.3b. As indicated, the UE with the highest data rate demands will pay more compared to the other clients in the system, at least for the two of the three cases. When applying the 3rd Policy, we observe that the UEs with the highest demands (UE2 and UE3) will pay less than the next UE in order (UE1). Due to this observation, we present in Figure 3.3c the normalized cost per Mbps for each UE for each policy. Figure 3.3c confirms that the 3rd Policy is not fair in terms of cost. UE4 with the lowest data rate demand will be charged the highest amount per Mbps.

Finally, we seek to investigate the impact of the ordering policy in the utilization of the available RATs of the system. For this purpose, we visualize, in terms of percentages, each UE's usage of each RAT for each one of the three policies as shown in Figure 3.4. We monitor that some UEs are affected more than others in the allocation of their TCs to the available RATs, based on the order that they will be probed to determine changes (or not) by the network controller. For instance UE0 and UE3 are following the same pattern through the different ordering policies, with almost the same distribution for the 1st and 3rd policy (Fig. 3.4a, 3.4c) and greatly differ for the 2nd (Fig. 3.4b). On the other hand, UE1 and UE2 are keeping similar distribution behavior across the three policies.

3.6 Chapter Conclusion

In this chapter, we proposed a dynamic pricing scheme for network selection in a Multi-RAT environment where the end-users have multiple applications concurrently running with diverse network requirements for each one of them. We consider mulit-homed UEs, meaning that they are able to use multiple access technologies simultaneously. These factors add up to the complexity of the resource allocation problem in such environments. For the evaluation of our scheme, we ported the

system model into a real-world setup, employing the NITOS testbed. Our experiments consider a heterogeneous network with multiple RATs (WiFi, 3G, 4G, WiGig) and multi-homed clients. The obtained results pinpoint the importance of the UEs' ordering policy that it is applied during the updating process of the system and how it may affect the cost incurred at each UE as well as the utilization of the available RATs. We proposed and examined three different ordering policies which are based on data rate demands, spectrum efficiency, and sensitivity. In the following chapter, we study the fairness of the aforementioned ordering policies, proposing a solution with dynamic selection among these policies during the execution of the experiments.

Chapter 4

Dynamic RAT Selection in 5G HetNets

Contents

4.1 Chapter Introduction

Ultra-dense heterogeneous networks (UDHetNets) are expected to play a key role during the transition from legacy to next-generation mobile networks. Through the dense-deployment of multiple access networks, the operators can take advantage of the added capacity of either smaller scale access points (e.g. femto-/pico-cells) and non-3GPP technologies like WiFi, towards offering enhanced capacity and lower latency services to the connected User Equipment (UE) [25]. At the same time, through the utilization of new wireless spectrum, available at the centimeter and millimeter Wave (mmWave) bands for fixed wireless access, and the enhanced spectral efficiency through massive MIMO techniques [30], ultra-high speed broadband services can be offered. The above technologies are widely considered as the main enablers for the 5th generation of mobile networking, annotated as 5G communications. Nevertheless, these technologies induce increased costs for the operators: from the mobile network operator's (MNO) perspective, induced CAPEX and OPEX costs on upgrading and maintaining the network respectively are higher, whereas the Average Revenue per User (ARPU) is slowly decreasing [76]. Therefore, solutions that do not disrupt the cost structure of operators need to be applied towards keeping track with the constantly rising user demands and the cellular network evolution.

Heterogeneous Networks (HetNets) can assist towards offering enhanced network capacity, through the densification of the offered network infrastructure. HetNets enable the operation of smaller scale dense heterogeneous access points (through

pico-/femto-cells), with existing or legacy technologies (e.g. NR, LTE, HSPA, UMTS, WiFi), complementing the macro-cell base station. Factors such as increased heterogeneity in the Radio Access Technologies (RATs), diversity in the coverage ranges and capacities of each cell, create a highly complex environment for selecting the appropriate combination of technologies per UE in order to maximize the delivered performance [111]. Multi-homing features at each network UE allow the concurrent operation of multiple wireless interfaces, and overall add up to the complexity [54]. Moreover, when considering the disaggregated nature of future base stations based on the Cloud-RAN concept [5], multiple technologies can be integrated at the base station level controlled by the operator [71], similar to legacy technologies like LTE WLAN Aggregation Adaptation Protocol (LWAAP) [4]. These technologies create fertile ground for the development of intercommunication mechanisms between HetNets, towards defining the best selection of technologies, in a per-packet basis, through which each user of the network will be served.

The variety of access options in the 5G concept comes with several mobile applications, characterized by different network requirements that thrive around this ecosystem (e.g. VR/AR, e-Health, etc.). Different applications may present diverse QoS requirements, depending on the class of their traffic (e.g. email applications vs real-time video streaming). Four types of traffic classes have been distinguished in the literature to describe QoS requirements for different data types: conversational real-time class such as voice traffic; interactive class such as web browsing; streaming real-time class such as video streaming; and background class (best-effort) such as emails [42].

Modern mobile devices are equipped with a plethora of network interfaces. With the introduction of multi-homing protocols such as Multipath TCP [97] and Stream Control Transmission Protocol (SCTP) [96], the concurrent use of multiple paths over multiple networking interfaces associated with different RATs is now a reality. Yet, considering that different access technologies may induce different charges for the mobile network users, the selection of the available RATs shall be restricted based on Utility Functions that consider the cost per each access technology. In this chapter we consider such environments, trying to answer the following key questions:

1. How pricing and congestion drive the decisions of UE's to direct their traffic classes through available RATs within an area.

2. How dynamic decisions by the network controller define the policy that drives the distributed decisions of served UEs.

3. How the cost per RAT is updated, subject to the access decisions that UEs take.

Towards addressing these questions, we are extending the contributions presented in Chapters 2 & 3. We assume multiple traffic classes per each network UE, with each class being able to use different RATs with different congestion levels and prices. The distributed decisions create an allocation problem with high complexity. Moreover, we introduce novel pricing and allocation mechanisms, and compare all of them under a unified set of experiments in real infrastructure. The solution is evaluated using new metrics, towards validating the overall network utilization efficiency. We assess the overall performance of our extended PMP scheme that employs dynamic pricing per each access class, using a realistic urban mobility model applied in a real-world testbed setup. We further develop our scheme and present two approaches, based on centralized versus fully distributed allocation decisions.

The rest of this chapter is organized as follows. In Section 4.2 we present a literature overview of heterogeneous network access schemes. In Section 4.3 we introduce

our system model. In Section 4.4 we provide our proposed distributed allocation system and the centralized reference model. Section 4.5 presents our system architecture and Section 4.6 includes our evaluation results on the performance of the proposed scheme for different polling and pricing policies. Finally, in Section 4.7 we conclude and present some future directions.

4.2 Related Work

A significant volume of research approaches has been published regarding congestion-based pricing in many different utility markets, such as transportation and electricity distribution networks [101]. In 1997, a first form of dynamic congestion pricing for the Internet was proposed in [70]. Since then, congestion pricing has been extensively studied in communication networks.

Nevertheless, the application of increased network heterogeneity for combating the capacity scarcity problem redefines the pricing problem. This can become even more complex if we consider that each user has multiple traffic classes, which can be simultaneously served by different access networks. In [112], an extensive categorization of different policies for network selection is provided, based on the utility functions and applied mathematical models. Similarly, authors in [69] provide an extensive survey for indicative applications of economic and pricing theory for the problems of user association in HetNets, spectrum allocation, interference, power management and wireless caching.

In [37], authors approach the network selection problem with the target to jointly optimize the utility functions of both the operator and the users of the network. Their evaluation denotes that network topology, user profile, user applications as well as type of sent traffic, highly influence the association problem. In [9], a network association mechanism is proposed when considering a D2D-enabled heterogeneous wireless environment. The authors develop a distributed solution leveraging user-centric network association in order to enhance system performance and support reliable connectivity.

Authors in [27] approach the network selection problem by using a rewards mechanism. Each user dynamically tunes the amount of traffic transmitted through each available access network based on rewards received by the network operator. The rewards reflect on the impact of each user's load to the cell. Contrary to this, in [8] a non-cooperative game formulation is provided for access selection, where each user of the network attempts to selfishly maximize its throughput. Based on the impact of a user's decisions on the rest of the cell's users, the existence of Nash equilibrium, to which the system converges, is investigated. As a means to evaluate this solution, the average number of per-user RAT switchings is monitored. Authors in [29] propose a traffic type-based solution for reputation-based network selection which makes a significant improvement in the system performance in terms of QoS for each traffic type.

In [58], a motivation mechanism is proposed, for encouraging external WiFi Access Points (APs) to provide network resources to the heterogeneous setup. The APs are provided with a salary and a bonus incentive in order to participate in data offloading, by offering one traffic class. The inclusion of such APs is formulated as a Stackelberg game. In [57], mobile users collaborate by forming mesh networks and share their Internet access. The authors use a virtual currency in order to create incentives and enable the user cooperation. In [75], the authors provide a pricing scheme for a Multi-RAT with LTE and WiFi technologies. In [118], pricing is applied

for the efficient coordination of spectrum resources, for the cooperation between small- and macro-cell providers. The problem of pricing-based dynamic service selection by mobile users was based on a Stackelberg differential game with open-loop Stackelberg equilibrium being the solution of this game. In [21], the user association problem in heterogeneous networks is considered as a joint problem with subchannel assignment and power allocation with the objective of maximizing the overall system throughput. A power adjustment algorithm for heterogeneous networks is also proposed in [102], assuming non-ideal communication links for downlink connections.

The authors in [85] apply the PMP algorithm in packet based networks (like the Internet), as a solution for service differentiation. They propose partitioning the network into several channels and use PMP as a tool for traffic diversion by applying different prices. In [18], authors investigate the applicability of PMP in a congested network environment, and analyze their contribution for different congestion functions. Both studies focus on the viability of the scheme in networking, and whether it may prove to be more profitable for the service provider, or achieve higher social welfare.

Leveraging the results and insights from these two works, we propose, apply and evaluate a dynamic pricing scheme based on PMP, with users taking decisions on the access technology allocation for each of their traffic classes based on current prices and congestion sensitivity. In [85] a flat-pricing based approach is proposed for congestion control on different logical channels. Our solution goes beyond the state of the art by introducing dynamic pricing for each access technology that is directly connected to congestion and user-induced traffic. In this context, UEs autonomously choose the allocation of their traffic classes to RATs, depending on how the channel conditions affect their utility maximisation objectives. In [18] it is investigated whether PMP can lead to social welfare for dynamic prices. It is proved that for certain congestion conditions and appropriate pricing social welfare is achieved. In this work we develop three policies and the dynamic selection of the appropriate policy under different congestion and channel conditions. The policy decision is made by the network controller, followed by the distributed RAT choices that UEs are making to route their traffic classes. The network controller decides centrally which policy will apply each time, with the objective to achieve distributively the closer performance to the optimal proportionally fair policy that would need to be centrally implemented.

In [51], the authors examine through a game-theoretic approach, the importance of the matching between pricing and service quality demands. They show that there is possibility for the system to settle into an unwanted equilibrium similar to the one produced in the "Prisoner's Dilemma" game. Moreover, the authors showcase that dynamic pricing approaches according to users' preferences over different service classes can lead to stable equilibrium.

This work is aligned with the proposed beyond 5G and next generation deployments, which consider a densely deployed Multi-RAT heterogeneous networking environment [109] and multiple traffic classes that pose new challenges on the implementation of aggregation mechanisms for augmenting the overall capacity and network coverage. We further examine and exemplify how different allocation decisions can widely affect the delivered performance to UEs, over a Multi-RAT environment. Some of the factors that have impact on the performance of the system include the policies applied at the operator side for selecting the sequence of requests that are examined and the RAT pricing schemes followed in each RAT. Moreover, we introduce a policy based on ordering the Multi-RAT UEs according to the Kendall tau

TABLE 4.1: List of notations and their physical meanings

Symbols	Physical Meanings
UE_i	UEs under the coverage of the Multi-RAT system
r_i^l	Data rate demand of UE_i's Traffic Class (TC) l
C_m	Total capacity of RAT m
θ_i^l	Sensitivity of UE_i's TC l
w_i	Normalized spectrum efficiency of UE_i, $w_i \in [0,1]$
Q_m	Mass of TCs allocated to RAT m
$U_i^l(m)$	Utility function of UE_i for TC l in RAT m
$\bar{C}_m$	Available capacity of RAT m
$f_m(Q_m, \bar{C}_m)$	Perceived congestion fucntion for RAT m
p_m	Access price of RAT m
c_i	Weighted prop. fair bandwidth allocation to UE_i
τ_j	Ranking of Policy j
τ_{pf}	Ranking of proportionally fair reference policy
τ_{pmp}	Ranking of the proposed dynamic PMP policy
$K(\tau_{pf}, \tau_j)$	Kendall's tau correlation coefficient

distance from the weighted proportionally fair ranking, that is optimal, and apply a dynamic scheme for selecting RATs per each traffic class per each client.

4.3 System Model

The main notations used in this chapter are summarized and explained in Table 4.1 for the ease of reading. We focus on a heterogeneous access environment of a cellular service provider, with M classes of available RATs. Each RAT class m refers to a specific access technology (e.g. WiFi, 3G, 4G, 5G-NR) of capacity C_m. Hence, the total available system capacity is equal to $C = \sum_{m=1}^{M} C_m$. Each UE_i is able to maintain L types of traffic classes, with sensitivity represented by $\boldsymbol{\theta}_i = \{\theta_i^1, ..., \theta_i^L\}$, with $\theta_i^l = r_i^l / w_i$ representing the sensitivity of the traffic class l, which is affected by the data-rate demand r_i^l and the normalized spectrum efficiency w_i of UE_i, with $w_i \in (0,1]$ for $i = (1, ..., N)$. This enables a UE to assess the channel quality in each traffic allocation decision, as it affects its payoff in the utility function of each traffic class it maintains. We denote as Q_m the mass of traffic classes served through RAT m, and we assume a normalised total mass of traffic classes equal to 1. Consequently, the mass of not allocated traffic classes in the HetNet is $Q_0 = 1 - \sum_{m=1}^{M} Q_m$.

We focus on the UEs that are able to associate with all the provided radio access technologies in the Multi-RAT environment. For each UE in the HetNet, a flat-rate charging V applies for accessing the system. Each UE maintains a congestion perception according to the function $f_m(Q_m, \bar{C}_m)$ that is affected by the mass of traffic classes served by RAT m and the available RAT's capacity denoted by $\bar{C}_m$. A price equal to p_m is charged to a UE that decides to allocate a traffic class (TC) to RAT m. Each UE makes distributed decisions to allocate a TC l, to the RAT m that maximizes its utility:

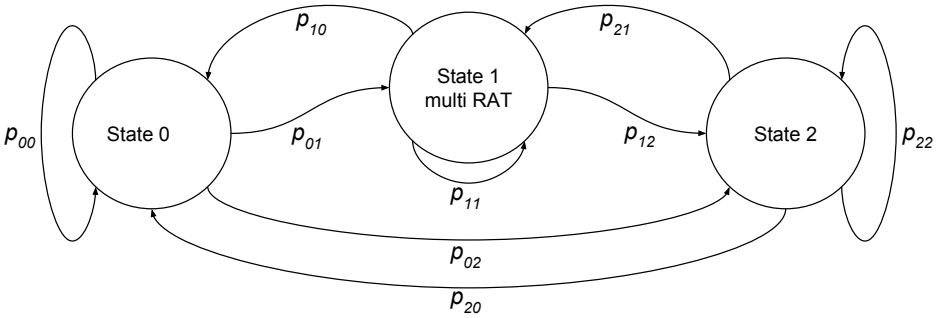


FIGURE 4.1: Markov Model States for the mobility of a UE

$$U_i^l(m) = V - p_m - \theta_i^l f_m(Q_m, \bar{C}_m) \tag{4.1}$$

such that:

$$m(\theta_i^l) = \underset{1 \le m \le M}{\operatorname{argmax}}\, U_i^l(m) \tag{4.2}$$

or no service if $U_i^l(m) = 0, \forall\ m \in (1, ..., M)$ and $\forall\ l \in (1, ..., L)$. The function representing the perception of the UEs for the congestion conditions in each RAT m is equal to:

$$f_m(Q_m, \bar{C}_m) = \frac{Q_m}{\bar{C}_m / C_m} \tag{4.3}$$

The congestion perception function allows the UEs to make decisions taking into consideration the available capacity of each RAT, as the value of eq. (4.3) affects the payoff of the utility function of each traffic class that is indicated in eq. (4.1)

As the price p_m for each RAT is decided by the HetNet network controller and then the UEs respond by allocating their traffic classes to maximize the utility of each of their traffic classes, a two-level (Stackelberg) game appropriately models this set of interactivity. In [87], we provide the Stackelberg game formulation that leads to Wardrop equilibrium [113], where no UE is having better payoffs by changing its traffic allocation decisions.

We approach the price setting by the HetNet network controller with two dynamic pricing schemes. The first is based on the available capacity of each access class m and it is common for every UE in the system. The second is based on personalized pricing (first-degree price discrimination), where each UE_i is charged for its traffic classes' throughput demands. In the first pricing scheme, the maximum price for each class $p_m^{\max}$, is set for accessing class m when its total capacity C_m is allocated, and the minimum price $p_m^{\min}$ is set when the total capacity of class m is available. The price as a function of available capacity $\bar{C}_m$ is expressed in (4.4)

$$p_m = \max\left(p_m^{\min}, p_m^{\max}\left(1 - \frac{\bar{C}_m}{C_m}\right)\right) \tag{4.4}$$

The second pricing scheme that we employ is based on the UE_i's TC throughput demands. We utilize two different functions for the price calculation and we define two rounds of pricing. In the first round we keep the same pricing function as before, stated in (4.4). In the second round, if part of the available RAT's capacity is already in use, the operator calculates the total pricing and used capacity of each RAT and then each UE_i is charged based on the ratio of used to total bandwidth in each RAT,

multiplied by its data rate demand. The price as a function of a UEs' throughput demand and the available RAT capacity $\bar{C}_m$ is expressed in (4.5)

$$p_m^{i,l} = p_m \left(\frac{C_m - \bar{C}_m}{C_m} \right) r_i^l \tag{4.5}$$

where r_i^l is the data rate demand of traffic class l of UE$_i$, $i = (1, ..., N)$ and $m = (1, ..., M)$.

Based on [81], we define the mobility of the UEs in our system which is presented in Figure 4.1. In this Markov model, the UEs that are in State 0 have 3 available options: to stay in this State with probability p_{00}, to enter the Multi-RAT system (State 1) with probability p_{01}, and to bypass State 1 and go directly to State 2 (p_{02}). The options for the other two States of the Markov model are defined accordingly. Next, we present our dynamic PMP scheme and we evaluate it versus a centralized solution, where the centralized network controller has the total information regarding the capacity needs of all traffic classes of the UEs in the HetNet. The bandwidth allocation in the centralized solution is done based on proportional fairness weighted by the average traffic sensitivity of each UE in the HetNet.

4.4 Distributed vs Centralized Allocation Decisions

In this section we describe two different approaches on the allocation of the system's resources: 1) a centralized solution, leveraging the weighted proportionality fairness concept, and 2) a distributed solution based on PMP allocation. We use the weighted proportionally fair approach for benchmarking reasons, so that the decision on the distributed dynamic policy presents the closest performance to the centralized solution. We hereby present the under-study system, the centralized allocation system and the dynamic distributed allocation system.

4.4.1 Centralized Allocation System

In this subsection, we describe the centralized solution in which the under-study system is presented as a utility maximization problem and provide the optimal solution for proportionally fair allocation of bandwidth resources among the UEs. Subsequently, based on the bandwidth allocation of the UEs we organize them in descending order. We consider that each UE in our HetNet has separate bandwidth needs for each of its l traffic classes (TCs), where $l \leq L$. We also assume that each TC has a sensitivity noted as θ_i^l, and the average sensitivity of the ith UE is $\overline{\theta}_i = \sum_{l=1}^{L} \theta_i^l / (\mathbf{1} \cdot \mathbf{d}_i)$, where $\mathbf{d}_i = \{d_i^1, ..., d_i^L\}$ and

$$d_i^l = \begin{cases} 1 & \text{if UE}_i \text{ has traffic class of type } l \\ 0 & \text{if UE}_i \text{ has no traffic class of type } l \end{cases} \tag{4.6}$$

If N UEs access the available RATs of the system, the bandwidth allocation for the ith UE can be written as c_i, $i = (1, ..., N)$, and $\sum_{i=1}^{N} c_i \leq C$, with C being the maximum capacity of the provided HetNets.

The proportionally fair solution, as shown in [62], is represented by an allocation vector $\mathbf{c} = (c_1, ..., c_N)$, which is (i) feasible, i.e. $\mathbf{c} \geq 0$ and $\sum_{i=1}^{N} c_i \leq C$ and (ii) is negative or zero for any other feasible vector $\mathbf{c}^*$. For average traffic sensitivities denoted as $\overline{\theta}_i$, a proportionally fair allocation shall satisfy the following:

$$\sum_{i=1}^{N} \overline{\theta}_i \frac{c_i^* - c_i}{c_i} \leq 0 \tag{4.7}$$

which can be rewritten as:

$$\sum_{i=1}^{N} \overline{\theta}_i (log(c_i))' dc_i \leq 0 \tag{4.8}$$

From eq. (4.8) we can infer that $U_i(\mathbf{c}) = \sum_{i=1}^{N} \overline{\theta}_i (log(c_i))$ is maximized with the allocation that follows the proportionally fair solution, which accrues from the solution of the following maximisation problem

$$\begin{aligned} \max_{\mathbf{c}} \quad & \sum_{i=1}^{N} \overline{\theta}_i log(c_i) \\ \text{subject to} \quad & \sum_{i=1}^{N} c_i \leqslant C \\ \text{and} \quad & c_i \geqslant 0, \forall i = 1, ..., N \end{aligned} \tag{4.9}$$

The problem in (4.9) has a unique solution since the utility function is strictly concave and the constraint set is convex. We relax the constraints and define the Lagrangian [15], and we change $c_i \geqslant 0$ to $-c_i \leqslant 0$

$$L(\mathbf{c}, \mu) = \sum_{i=1}^{N} \overline{\theta}_i (log(c_i)) - \mu_0 \left(\sum_{i=1}^{N} c_i - C \right) + \sum_{i=1}^{N} \mu_i c_i \tag{4.10}$$

with $\mu_0 \geqslant 0$ and $\mu_i \geqslant 0, i = 1, ..., N$. Next, we start with the stationarity condition of the Karush-Kuhn-Tucker (KKT) optimality conditions and we have

$$\nabla_{c_i} L(\mathbf{c}, \mu) = \frac{\overline{\theta}_i}{c_i} - \mu_0 + \mu_i = 0 \tag{4.11}$$

and since $\overline{\theta}_i > 0$, we have that $\mu_0 > \mu_i$, which means that $\mu_0 > 0$. Taking the complementary slackness conditions we have

$$\mu_0 \left(C - \sum_{i=1}^{N} c_i \right) = 0 \tag{4.12}$$

$$\mu_i c_i = 0 \tag{4.13}$$

$$\mu_0 \geqslant 0 \text{ and } \mu_i \geqslant 0, i = 1, ..., N \tag{4.14}$$

As $\mu_0 > 0$, it follows from (4.12) that

$$\sum_{i=1}^{N} c_i = C \tag{4.15}$$

which means that c_i, $i = 1, ..., N$ cannot be zero. Therefore by forcing $\mu_i = 0, \forall i = 1, ...N$ we have from (4.11)

$$c_i = \frac{\overline{\theta}_i}{\mu_0} \tag{4.16}$$

By combining (4.15) and (4.16) we reach the optimal solution which represents the weighted proportionally fair solution

$$c_i = \frac{\overline{\theta}_i}{\sum_{i=1}^{N} \overline{\theta}_i} C \tag{4.17}$$

When the UEs are organized in descending order, with their allocated bandwidth being the ordering keys, we extract the proportionally fair ranking denoted by τ_{pf}. In the next section, we provide different policies for examining the allocation per each UE, in order to provide a distributed solution to the allocation decisions. We use different ranking parameters for each of the under study policies, and develop a dynamic policy that selects and applies, in every system cycle, the closest ranking policy to τ_{pf}. We mark as τ_{pmp} the ranking based on our dynamic PMP policy, and we use Kendall's correlation coefficient $K(\tau_{pf}, \tau_{pmp})$ in order to determine the ranking distance between the dynamic distributed scheme and the optimal centralized solution. The Kendall tau distance measures the pairwise disagreements between ranking lists consisting of the same set of elements [43].

TABLE 4.2: Ranking policies for polling the UEs of the HetNet

Policies	Policy Symbol	Ranking Parameter	Description
Policy 1	τ_1	$\sum_{l=1}^{L} r_i^l$	Descending order of aggregate data rate demands of the UEs
Policy 2	τ_2	w_i	Descending order of normalized spectrum efficiency of the UEs
Policy 3	τ_3	$\overline{\theta}_i$	Descending order of the average sensitivity of UEs traffic classes
Policy 4	τ_{pf}	c_i	Descending ranking of the proportionally fair bandwidth allocation (Reference)
Policy 5	τ_{pmp}	$\max(K(\tau_{pf}, \tau_j))$	For $j = (1,2,3)$: Dynamic selection between Policies 1-3

4.4.2 Dynamic Distributed Allocation System

Considering the second approach, the procedure of the distributed allocation of TCs to RATs is invoked in two different manners: 1) periodically, for all the UEs being served from the Multi-RAT HetNet, and 2) each time that a UE enters or exits the coverage area of the HetNet. We consider such events as cycles of the distributed system (system cycles), during which the prices per each access technology are updated and each served UE is polled in sequence. In the cases that these events occur, each polled UE, either already being served or entering the HetNet, decides distributively its resource requests based on the updated system report. As shown in [88], the sequential polling creates fertile ground for applying different policies with respect to the manner with which we order the UEs participating in our network, and is used for each UE to announce their requests in bandwidth. In Table 4.2 we

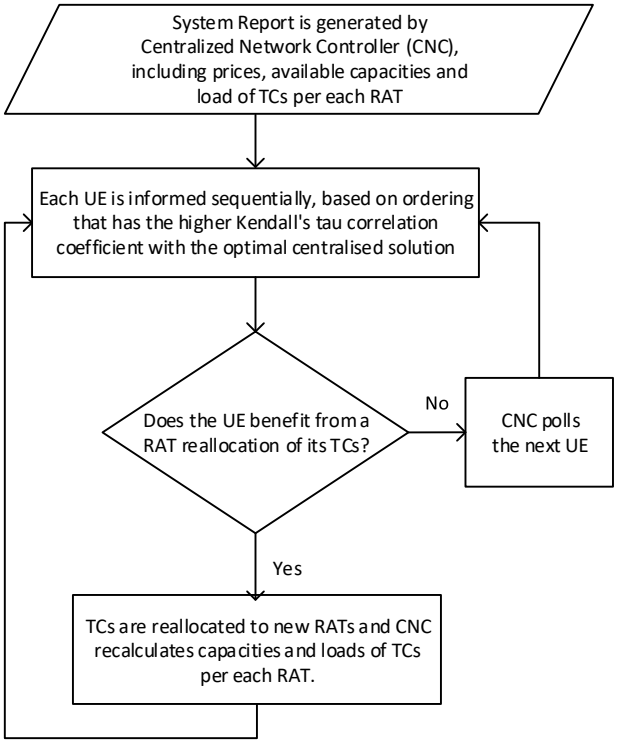


FIGURE 4.2: Centralized Network controller's polling functionality flowchart during a system cycle

showcase the policies that we apply to our system, along with the polling order and reference ranking that each one applies. In this work, we employ a dynamic selection of the ordering policy noted as τ_{pmp}. This policy dynamically achieves the closest ordering to the weighted proportionally fair approach (optimal centralized solution), in a completely distributed manner.

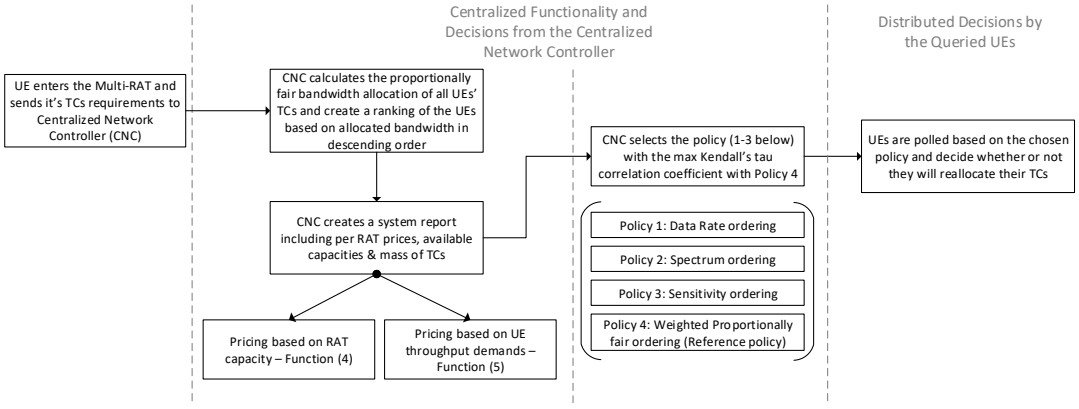


FIGURE 4.3: Schematic representation of the system model when a UE enters/exits the multi-RAT system

During each system cycle, we calculate the correlation coefficients $K(\tau_{pf}, \tau_j)$ for $j = (1,2,3)$ based on the calculated congestion per each RAT of the HetNet, and choose the policy with the maximum Kendall's tau correlation coefficient to τ_{pf}. In Figure 4.2, we present the steps taken for the reallocation of the TCs of each UE based on the chosen dynamic polling policy. The pricing framework is running at

the Central Unit (CU) entity of a disaggregated Multi-RAT base station, as it is the closest to the edge convergence point for all the different RATs belonging to the same provider, integrated in a software entity that we call Centralized Network Controller (CNC). In each cycle of the system, all UEs are polled according to the chosen policy and are queried on whether they prefer to reallocate their TCs to other RATs or not. The decisions that are made by each UE are based on: 1) the current price for accessing the RAT under consideration, 2) the sensitivity of the examined TC and 3) the perception of the UE for the congestion conditions of the RAT. Each query is initiated by a system report that is sent to polled UEs, containing RAT prices, available capacities and load of TCs per each RAT. Based on the system report and the utility function that corresponds to each TC, the UEs are enabled to take distributed decisions on the reallocation of their TCs to another RAT. Their objective is to maximize their corresponding utility functions, shown in eq. (2), for each served TC. After all the UEs are queried, CNC consecutively queries the new UEs that entered the HetNet. If changes are identified, the rankings are recalculated and CNC checks whether another ordering policy shall be applied, based on the Kendall's tau correlation coefficient between the current and target policies, as shown in Figure 4.3. The figure illustrates how the policy decisions are dynamically concluded at the network controller by selecting the policy with the higher Kendall's tau correlation coefficient with the proportionally fair ranking policy. More precisely, each time a system event occurs, the CNC calculates the proportionally fair bandwidth allocation of all UEs for benchmarking purposes and the prices based on the applied pricing scheme (either based on RAT capacity or UE throughput demand). The next step for the CNC is to select the ordering policy comparing Policies 1-3 to the benchmarking policy (Centralized Allocation System). Following this, the distributed decisions of the queried UEs are made, with the objective to maximize the utility functions that correspond to their served traffic classes. In the next section, we describe the system architecture under which we evaluated our scheme, and the developed algorithms for the distributed allocation of the TCs to multiple RATs.

4.5 System Architecture

The proposed scheme is considered for four different radio access technologies: 1) 3G RAT, 2) WiFi RAT, 3) 4G RAT and 4) 5G RAT. Towards validating our scheme in a real life environment, we measured the capacity of each RAT in the NITOS wireless testbed [83], a wireless network ecosystem that provides access to these technologies for experimentation, and took them into consideration in the model's evaluation. The measured capacities and setups are presented in Table 4.3. These are reference values that the equipment has been configured to work with, as they are commonly found in commercial off-the-shelf (COTS) equipment. For the WiFi settings, these

TABLE 4.3: Configuration and measured capacities for the RATs under consideration

RAT	Capacity	Configuration
3G	42 Mbps	HSDPA 5MHz - 2x2 MIMO
WiFi	130 Mbps	802.11n 20MHz - 2x2 MIMO
4G	70 Mbps	LTE 10MHz - 2x2 MIMO
WiGig	700 Mbps	802.11ad 1.76GHz - 12x1 Ant. Elem.
5G	692 Mbps	NR FR2 (24250 MHz – 52600 MHz), 400MHz, 2x2 MIMO TDD

are the maximum goodput values that a single channel WiFi access point can work with for a two antenna configuration. For the 4G and WiGig settings, these are the maximum values that the COTS equipment in the testbed can be configured with. As there is currently no 5G-NR RAT available in the testbed, we performed our testbed experiments using two WiGig nodes, operating in the 60 GHz band. Since our algorithms consider only the performance characteristics of each technology, like the network capacity, and not the low level physical layer settings (e.g. MIMO settings, number of antennas, etc.), we believe that no deviation exists when running the experiments in such settings. As a matter of fact, the testbed where we conduct our experiments is RF-isolated, and thus there is no external interference that could potentially impact our results. Our experiments consist of both simulations and real-life experiments. The topology that we use for evaluating our contributions is shown in Figure 4.4. We consider a heterogeneous disaggregated base station setup, following our contributions in [71]. The higher layers of the stack are running as services in the Centralized Unit (CU) of the stack, whereas multi-technology network access is realized through heterogeneous Distributed Units (DUs) offering 3G, 4G, WiFi and WiGig connectivity. In the following paragraphs we detail the different modules developed for evaluating our contributions.

We develop two algorithms for implementing our system model, placed at the network edge inside the CU side of a disaggregated base station or completely distributed, running at the UEs. The system is initialized by the controller selecting in a uniformly random manner the TCs of State 1 network UEs to the available access technologies. Following this, the CNC determines the capacity, prices and load of TCs per each technology. Subsequently, the controller transmits an announcement to all the UEs of the network, based on the selected ordering of users as noted in the previous section, and listens for any replies. If a UE enters State 1, the CNC approves or denies access to the system, after calculating whether it can be served by the remaining capacity. Every change in the HetNet (UEs reaching/leaving State 1) triggers an update in all the values of the system and subsequent announcements to all the UEs in State 1, which further calculate their utility functions. As we mentioned, the policy for delivering the system announcements is highly dependant on the manner through which we organize the UEs. The applied ordering is based on network rate demands, sensitivity, spectrum allocation, or a dynamic weighted selection of them. The algorithms running at the UE side (distributed) are communicating with the CNC. In the case that a UE enters our Multi-RAT system, it receives an announcement message that lists the remaining capacity of the network, the price per RAT, and the current load of all the TCs per each RAT. Subsequently, the utility per each TC of each UE is calculated, based on the available technologies, as shown in (4.1). Further on, each UE informs CNC on the RAT allocations it has concluded for its TCs. While in the Multi-RAT system, each UE periodically receives the announcements on the network utilization sent by the CNC. In such cases, and given that the UE remains in State 1, its utilities are recalculated and its TCs are reallocated to new RATs.

At each system cycle we execute the dynamic selection of ordering policies, and from that point we query the connected UEs for the status of their TCs and the preferred technology per each TC. We use 4 TCs with limits in the lowest and highest data throughput that they request. Each UE uniformly selects the throughput value from the ranges provided in Table 4.4. The represented values are based on [105] and we assume that each UE is able to maintain concurrently up to 4 different TCs.

In the Markov model (Figure 4.1) for the mobility of a UE, we configure the probabilities of State 0 with p_{01} to be equal to 0.5, the probability p_{02} to 0.2 and thus the

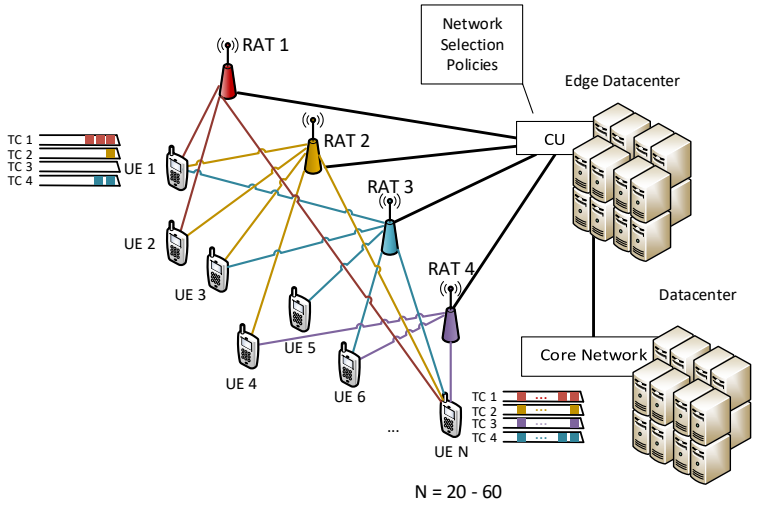


FIGURE 4.4: Overall System Architecture for the evaluation of the proposed resource allocation scheme

probability of p_{00} is 0.3. For State 1, the probabilities are defined as follows: p_{11} is configured to 0.5 and p_{12} to 0.45, meaning that p_{10} equals 0.05. As for State 2, we set the probabilities p_{20} and p_{21} equal to 0.05 and the probability p_{22} equal to 0.9.

TABLE 4.4: Traffic Classes specifications

TC Identifier	**Throughput Range (Mbps)**
TC1 - Background	0.3 - 1.0 Mbps
TC2 - Interactive	1.1 - 2.5 Mbps
TC3 - Streaming	2.6 - 8.0 Mbps
TC4 - Conversational	8.0 - 25.0 Mbps

4.6 Evaluation

We evaluated our proposed system in a two-fold approach:

1. We employed simulations aiming to investivate its capabilities in a large scale environment.

2. We ported its functionality to NITOS testbed with the objective to experimentally verify its performance in real-life wireless conditions.

We employed a Python-based simulator for the former and we integrated it with the disaggregated base station network implementation, consisted of a Centralized Unit and multiple heterogeneous Distributed Units, as developed in [71] for the latter. We organized five different experiments for each evaluation category and we hereby pinpoint the insights of the obtained results. Throughout the execution of the experiment, the CNC recalculates the RATs' prices and announces them to the UEs in the system. The proposed system is evaluated for its performance in terms of efficient bandwidth allocation among the available RATs, RATs price variations, fairness of the clients' ordering and induced cost to clients served by the HetNet.

For our simulations, we repeatedly evaluated our model for 10, 20 and 40 users, who were initially placed in State 1 and 40 users in State 0. We examined five different policies for the UE ordering: In Policy 1, the UEs are ordered based on their total

data rate demand. In Policy 2, the ordering is based on the normalized spectrum efficiency that each UE experiences using the wireless channel of each RAT. In Policy 3, the ordering of the UEs depends on their perception of congestion (sensitivity) in the available RATs of the system. Policy 4 is the reference policy for this evaluation. It solely runs in the CNC and sorts the UEs based on the weighted proportionally fair bandwidth allocation that is re-evaluated in every system cycle. Finally, in Policy 5 the Kendall's tau correlation coefficient [43] between policies 1-3 and policy 4 is calculated, and the policy (1, 2, 3) with the minimum calculated distance from the optimal policy 4 (higher correlation coefficient), is selected for every system cycle.

4.6.1 Extensive simulations

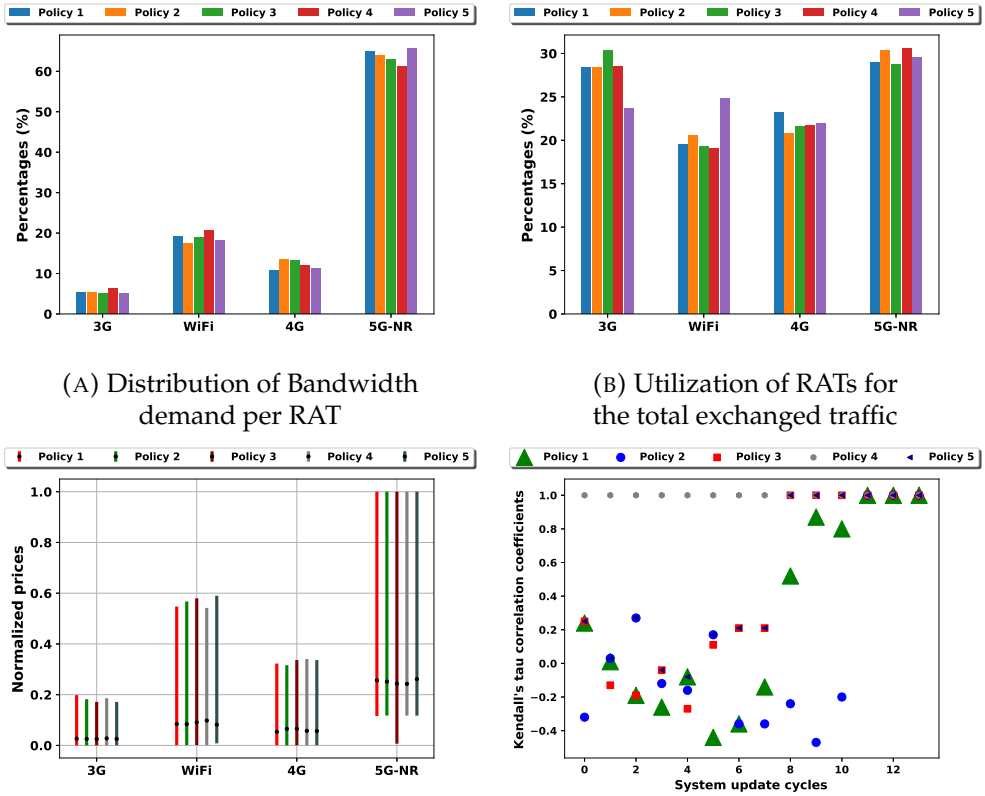
In this section, we decided to present the set of experiments where 10 *UEs* are placed in State 1 and 40 *UEs* in State 0, since it's closer to the testbed setup, allowing us to compare them more accurately. We organize our simulations in two categories based on the applied pricing scheme. First, we assess the proposed scheme utilizing the pricing function (4.4) and then we evaluate the system utilizing the pricing described by eq. (4.5).

Next we present the scenario under which we evaluated our proposed dynamic policy selection algorithm for both pricing functions: We start by placing 10 UEs in the Multi-RAT system and connect randomly their TCs to the different available RATs of the HetNet (3G, WiFi, 4G, 5G-NR). During the experiment, new UEs may enter the system or the existing UEs may decide to alter their TCs' allocation over the available technologies or even leave from the system. More specifically, throughout the experiment, 30 *UEs* out of the 40 in State 0 will enter the system, 8 will choose to bypass the system and for another 2 *UEs* the access will be denied by the Centralized Network controller, as their needs cannot be satisfied by the system. This results to an experiment in which 50 UEs participate in our proposed scheme's evaluation.

Pricing Based on RAT Capacity

In this set of experiments we employ the pricing function (4.4) for calculating the access price for each RAT of the system as we described in Section 4.3. In this case, the available capacity of each RAT is the parameter that forms the access price. To assess the proposed dynamic solution, we investigate the extend to which the system is affected in terms of RAT costs and allocated capacities by applying different ordering policies under the same scenario. Figure 4.5a shows how the overall bandwidth demands are distributed across the available RATs of the system and Figure 4.5b provides the total access technology utilization for the scenario under examination and for each one of the five polling policies.

As shown in Figure 4.5a, we conclude that for all the applied policies, over 60 % of the overall bandwidth is being traversed over the 5G-NR technology, used mainly for transporting TC4 traffic. As we mentioned earlier, each network UE is deciding the allocation of RATs per each TC based on the maximazation of its utility function (4.1) and the announced prices. These prices are published to UEs after all the allocations have been made and are depending on the remaining capacity of each technology. Based on the results presented in Figure 4.5b, we can detect a trend of equal utilization of the WiFi and 4G RATs and that throughout the evaluation of the different policies the percentages of utilization of each RAT remain in the same levels. In Figure 4.5c the variations of the allocated prices during our experiment are depicted. In order to directly compare all the under-study policies, we use normalized costs,



(A) Distribution of Bandwidth demand per RAT

(B) Utilization of RATs for the total exchanged traffic

(C) Normalized costs per each RAT

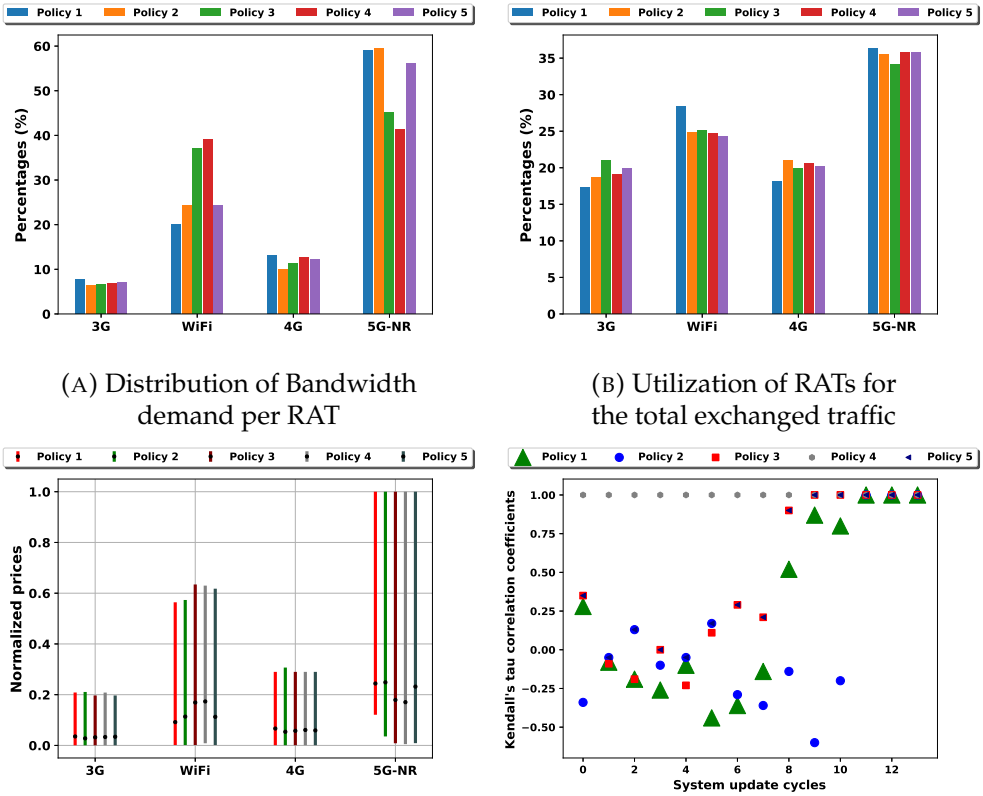
(D) Kendall's tau coefficient for the examined policies

FIGURE 4.5: Experiment results related to pricing policy based on RAT capacity

considering the maximum price achieved over the course of five experiments where each of the 5 policies was applied. The lowest and highest values per each RAT are shown in Figure 4.5c, whereas the average values are represented as dots. Based on these results, we can infer that Policy 3 and the dynamic Policy 5 outperform the rest when comparing with respect to price, as denoted in Figure 4.5c. Next, we wanted to compare the proposed Policies with Policy 4 in terms of fairness. Figure 4.5d presents the Kendall's tau correlation coefficient of each Policy with the optimal Policy 4. Following some initial iterations, we observe that policies 1 and 3 are very close to the optimal (Policy 4), whereas the 2nd Policy is the least fair.

Pricing Based on RAT Capacity and UE Throughput Demand

Based on the results from 4.6.1, we design and employ a different price scheme to our system and execute the same scenario as before in order to compare the two pricing schemes. As we already described in Section 4.3, in this pricing scheme there are two rounds of pricing instead of one. The algorithm in the Centralized Network Controller creates a system report for each UE as before, but in this case utilizing function (4.5). In this pricing scheme, we utilize two different functions for the price calculation and we define two rounds of pricing. In the first round we keep the same pricing function as before, stated in eq. (4.4). In the second round, the Centralized Network Controller calculates the total pricing and used capacity of each RAT and then each UE is charged based on the percentage of the total BW that it will use in each RAT under the current load conditions. Thus the parameters that form the access prices are the available capacity of each RAT and the UEs' TCs throughput



(A) Distribution of Bandwidth demand per RAT

(B) Utilization of RATs for the total exchanged traffic

(C) Normalized costs per each RAT

(D) Kendall's tau coefficient for the examined policies

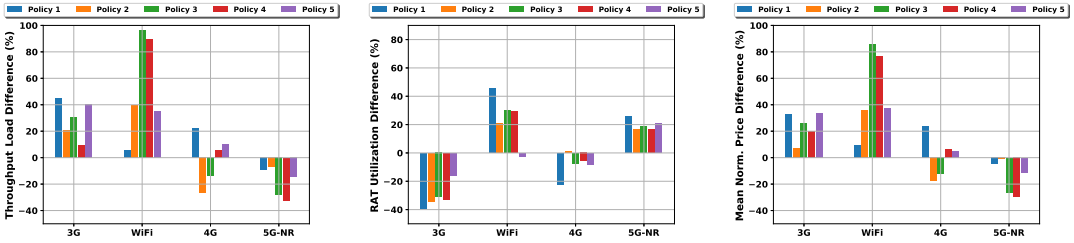
FIGURE 4.6: Experiment results related to pricing policy based on RAT capacity and UE throughput demand

demands. Then, according to the selected polling policy, it sends the report to each UE so as to decide whether or not it will reallocate its TCs to the available RATs of the system.

In Figure 4.6, we present the results of the new pricing scheme and we can infer that there are differences between the pricing schemes in the defined metrics except from the Kendall's tau coefficients which remain the same for both experiments as it is expected (Figure 4.6d). As for the distribution of bandwidth demand to the available RATs, we observe that the percentages of the 5G-NR RAT are decreased with this pricing scheme as shown in Figure 4.6a. In addition, we monitor a dependency between the WiFi and 5G-NR when we apply the Policies 3 and 4, in which the BW demands are reallocated from 5G-NR to the WiFi RAT. Figure 4.6b showcases the utilization of the available RATs of the HetNet for the total exchanged traffic during the experiment. Contrary to the results from 4.6.1, there is a decrease of approx. 10% on 3G which is reallocated to WiFi and 5G-NR RATs. Figure 4.6c showcases metrics of the allocated prices for each applied ordering policy during the experiment and we can observe an increase on the 3G and WiFi RAT and a decrease on the average values of the 5G-NR.

Moreover, we calculated the percentage of increase or decrease for each metric for all polling policies between the two pricing schemes and the results are depicted in Figure 4.7. The new pricing scheme steers the TCs of the UEs to the RATs with higher capacity (5G-NR and WiFi) and also mitigates the congestion by moving away excessive traffic from 4G RAT as it is shown in Figures 4.7a and 4.7b. In Figure 4.7b we observe that the mass of TCs is decreased on 3G and 4G RATs, choosing to be allocated in the remaining RATs of the HetNet. We monitor a trend on the 3G

RAT, where there is an increase of the bandwidth demands and simultaneously a drop on the number of TCs choosing to be served by this RAT. This means that TCs, on the 3G RAT, with lower throughput demands are choosing to be reallocated to WiFi and 5G-NR and less TCs with higher demands are utilizing the 3G RAT. Thus the new pricing scheme steers more TCs to RATs with high capacity and fewer TCs with higher bandwidth demands to RATs with lower capacities. As for the RAT access prices, we observe in Figure 4.7c that for the case of 4G and 5G-NR a decrease in prices takes place for most of the ordering policies. On the other hand, increases are detected on 3G and WiFi RATs. Based on the results in Figures 4.7a and 4.7b, these alterations in access prices are expected despite the increased utilization of 5G-NR RAT as the pricing scheme described in eq. (5) is highly depended by the throughput demands and in this RAT a decrease on its throughput load (Figure 4.7a) is monitored. Same behavior is noted in the 4G RAT for the polling policies 1, 4, 5 and as for the WiFi RAT, increases are monitored both in throughput load and its utilization.



(A) Impact on Throughput per each RAT (%)

(B) % Increase/Decrease in utilization of RATs for the total exchanged traffic

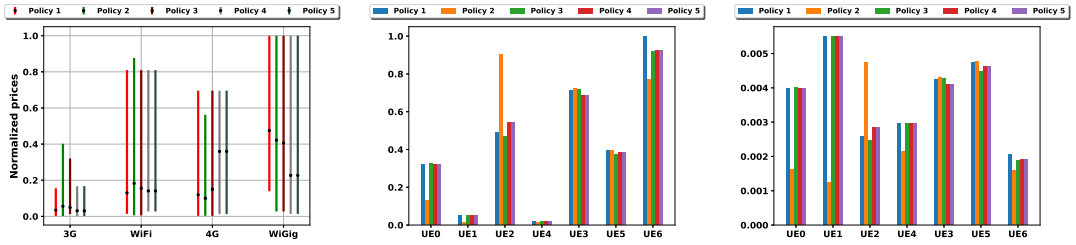
(C) % Increase/Decrease in normalized average price per RAT

FIGURE 4.7: Comparison between the two pricing schemes for the same scenario

4.6.2 Testbed experiments

With the aim to validate our simulation results, we execute our proposed scheme under a real world environment, utilizing the NITOS testbed. We determined a setup of clients in the testbed complying with the needs of our model and more specifically, we disabled the rate adaptation algorithms for all the technologies and set them up to the highest available configurations. Moreover, we logged RSRP and RSSI values for the LTE (4G) network equal to -76 dBm and -54 dBm respectively and similar values were observed for the WiFi and HSDPA (3G) networks in the Multi-RAT system as well. As for the testbed there is no 5G-NR technology available, so we performed our experiments using the WiGig technology which achieves similar performance in terms of capacity.

For this experiment, we employed 3 multi-homed UEs initially in State 1, another 5 UEs in State 0 and the scenario was formed as follows: The experiment starts with UE0, UE1 and UE2 in State 1, while UE3-7 will be entering the system later. UE7 decides not to enter the Multi-RAT system and selects the direct route to State 2 where it remains during the execution of the experiment. We ran the experiment for each of the five cases ten times. Due to the fact that the trend was similar to the one from the simulations, we chose to present only the results from the proposed scheme when utilizing the pricing scheme based on the UE throughput demands which is described in 4.6.1. In Table 4.5, the results of the total average achieved throughput for the UEs participating in the experiment are presented alongside with the number



(A) Normalized price variation per RAT

(B) Normalized cost per UE

(C) Normalized cost per Mbps

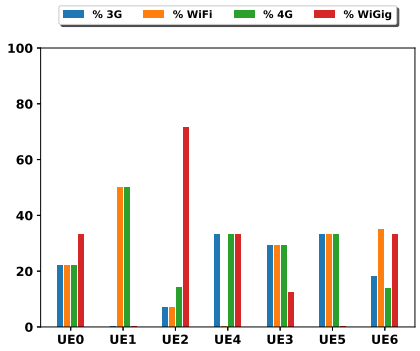
FIGURE 4.8: Testbed experimental evaluation of the proposed resource allocation scheme

of cycles remaining in the system. Table 4.5 showcases for each UE of the system, the achieved aggregated throughput of its TCs per system cycle. Furthermore, we observe that there are UEs in the system with different behavior and mobility. For example, UEs 0, 1, 4 remain in the system for a short number of system cycles while the others for at least 9 cycles and this affects their costs for accessing the HetNet.

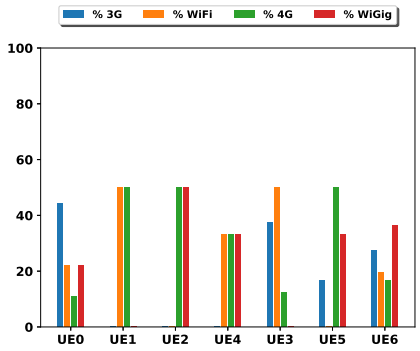
TABLE 4.5: Total Average Achieved Throughput per UE

UE ID	Throughput (Mbps)	# Cycles in the system
UE0	26.99	4
UE1	9.32	2
UE2	27.26	9
UE3	20.88	12
UE4	6.91	4
UE5	20.73	9
UE6	22.13	17

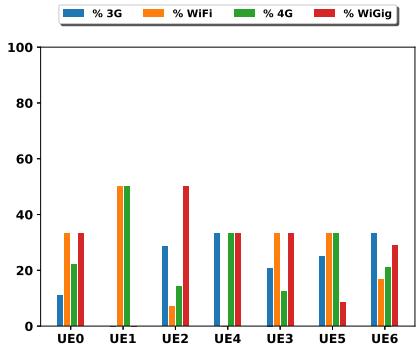
In Figure 4.8, we present the results from running the above scenario in the NITOS testbed and more specifically the normalized costs for accessing the Multi-RAT environment for each RAT and each UE of the system. Figure 4.8a verifies the simulation results in which the price values have the same trend as before with the WiGig having the highest and 3G the lowest. As shown in Figure 4.8b, UE2 and UE0 that have the highest demands among the UEs of the system, will be charged less than the UE6 which is next in the data rate demand ordering. Moreover, UE5 will pay less than UE2, even they both stay the same number of cycles in the system but UE5 has lower throughput demands. On the other side, UE3 that has lower demands will pay more due to the higher number of cycles in the Multi-RAT than UE2. This occurs due to the different types of TCs that those UEs had, the TCs allocation in the available RATs, and the number of cycles that each UE remained in the Multi-RAT system as presented in Table 4.5. This difference is the outcome of the personalised pricing that is applied to the system. Each UE with high data rate demands will contribute to a RAT's load with higher percentage when it allocate its TCs to this RAT. This explains why higher charges apply differently to some UEs compared to others, depending on the RAT's load status and the effective pricing scheme (eq. (5). Following this observation, in Figure 4.8c we provide the normalized access cost in relation to the achieved throughput for each UE and for each policy. UE6 with the third highest level of rate demands will be charged the lowest amount per Mbps due to the fact that it remained longer into the Multi-RAT system, and its requests are spread into multiple system cycles.



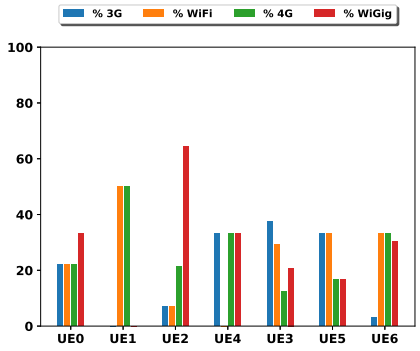
(A) 1st Policy - Data Rate ordering



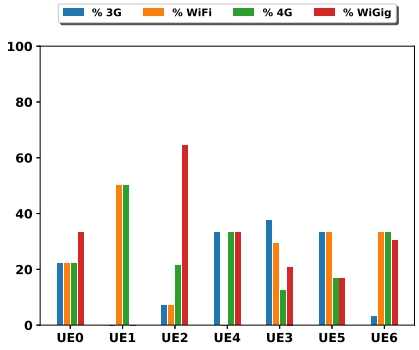
(B) 2nd Policy - Spectrum ordering



(C) 3rd Policy - Sensitivity ordering



(D) 4th Policy - Weighted Proportionally Fair bandwidth ordering



(E) 5th Policy - Dynamic selection between Policies 1-3

FIGURE 4.9: Data Rate distribution of each client across the System's RATs for the applied ordering policies

Finally, we evaluate the extend to which the level of RATs utilization is affected by the proposed ordering policies throughout the experiments. Towards this direction, we calculate the utilization of RATs for each UE's total throughput demands as it can be inferred from Table 4.5. The calculations are made for each one of the five policies under consideration, as it can be seen in Figure 4.9, and the results are in the form of percentages. We observe that some UEs throughout the different applied ordering policies, keep the same or similar allocation of their TCs to the RATs of the system, since others are highly affected resulting to altered percentages of RATs utilization for the same data rate requirements. For instance, UE3, UE5 and UE6 are following the same pattern for Policies 4 and 5 (Figures 4.9d and 4.9e) and greatly differ in the rest policies. On the other hand, UE1 keeps the same distribution among all policies, and UE4 across the four out of five policies. Furthermore, we observe

in Figure 4.9a where Policy 1 is applied, that UE0 and UE2 maintain the same RAT allocation with the experiments involving Policies 4 and 5. Based on Figure 4.9b, we can infer that Policy 2 affects the majority of UEs in the system, resulting in raises in RAT prices as it is also shown in Figure 4.8a. The results presented in Figure 4.9 provide fertile ground for monitoring the behavior of the UEs and further classify the UEs and apply different pricing schemes and ordering policies according to their demands towards achieving more effective utilization of the overall system. The main objective of the personalised pricing that we have introduced in this work is to shift users' preferences to RATs where they contribute as less as possible on increasing the existing traffic load. The necessary throughput for each traffic class in the system is satisfied, covering the corresponding QoS demands, without targeting to throughput increase but to congestion avoidance. As shown in Figure 4.7a and 4.7b, the evaluation of our new personalised pricing scheme reveals its impact on increasing the UEs' TCs allocation on 5G-NR and WiFi RATs and consequently their utilisation, mitigating the congestion that UEs in the 3G and 4G RATs would otherwise experience.

4.7 Discussion and Chapter Conclusion

In this chapter, we developed an Ultra Dense Heterogeneous system that decides on how the UEs shall select the RATs for each of their traffic classes to be served based on different policies and dynamic pricing of each RAT. We modeled and evaluated different policies, based on the UE's data rate demands, spectrum efficiency and sensitivity to network conditions and compared them to the optimal, proportionally fair allocation. The obtained results steered us to design and implement a new policy selection scheme that gives to the Centralized Network Controller the functionality to dynamically select, in each system cycle, the appropriate policy from a predefined policy set. In each cycle, the selected policy presents the closest possible performance to the reference policy of proportionally fair bandwidth allocation, identified through its Kendall tau correlation coefficient.

We provided detailed modeling of the pricing algorithms based on the Paris Metro Pricing scheme and the proportionally fair bandwidth allocation that was used as the reference policy. Initially, the pricing was based on the availability of access technologies' bandwidth and then it was further extended to include the rate demands of the UEs' traffic classes. We based our evaluation on simulations and testbed experimentation that enabled the assessment of the bandwidth allocation efficiency, the variation of access prices, the policy fairness and the induced cost to UEs covered by the HetNet system. The developed scheme is considering the optimization of the network operation as a whole, subject to each UE's congestion sensitivities. Therefore, the scheme is providing solutions regarding the stable operation of the entire under-study Multi-RAT.

In the future, we foresee to further extend our scheme by employing artificial intelligence classification algorithms that will enable UEs to further automate the access technology selection decisions aiming to minimize the induced access costs. Moreover, we will focus on the application of novel mobility schemes for the clients entering and leaving the Multi-RAT system and investigate how they affect the overall system.

Chapter 5

Employing MEC in 5G Heterogeneous Cloud-RANs

Contents

5.1 Chapter Introduction

The fifth generation of mobile networking (5G) is fostering several advancements in the access, edge and core network, promising to offer higher network capacity with lower latency, allowing a variety of (critical) services to thrive around this ecosystem. 5G also benefits from the wide application of Multi-access Edge Computing (MEC) [78], by serving network users directly from the network edge. Such functionality is highly beneficial especially for resource constrained devices such as mobile phones, as they can offload several parts of processing to the network edge. In the concept of Multi-access Edge Computing, heterogeneous technologies reside in the user access network, adding up to the overall network capacity by forming ultra-dense networks.

Along with the full integration of the MEC functionality in the network, 5G brings several new advances in the overall network architecture. The advent of Cloud-RAN design [73] for base stations allows the re-conception of technology solutions like MEC. 5G-New Radio (NR) specifications [1] define the disaggregation of the base stations based on the 3GPP Option-2 split [3], between the Packet Data Convergence Protocol (PDCP) and Radio Link Control (RLC) of the mobile stack. This creates two entities: the Central Unit (CU), located at the edge, and the Distributed Unit (DU), providing the actual access network. DUs can provide different access technologies (5G NR, 4G LTE or WiFi), served through the same CU. Each

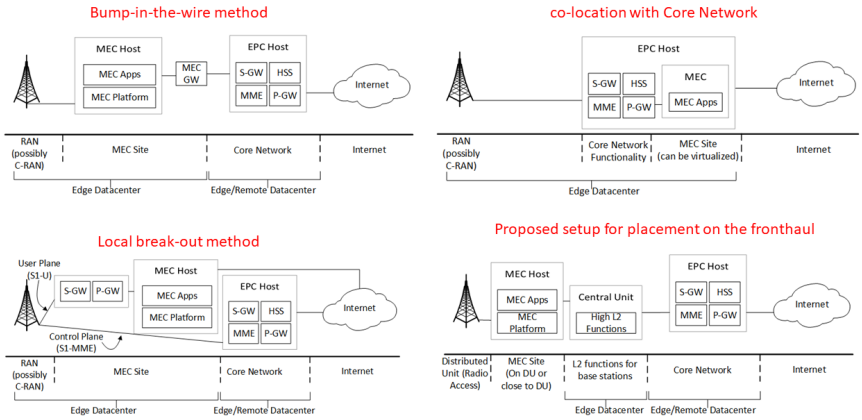


FIGURE 5.1: Existing vs proposed MEC service placement methods

CU may control simultaneously multiple DUs, allowing the aggregation of heterogeneous links for serving each network client. Although MEC has been considered as a low-latency enabler for 5G, its application is merely mapped to the new network architecture; its integration considers the same placements for the edge resources as in previous solutions (e.g. LTE) despite the new architecture.

In this chapter, we propose and experiment with the MEC service placement closer to the network edge, collocated with the DUs of a disaggregated base station setup, realizing truly the Edge Computing concepts. The services provided to the network users are placed at the base station fronthaul, thus minimizing access latency. Such setup is highly beneficial especially for the mobile users, as with the integration of Network Functions Virtualization (NFV) enablers, allows the real-time migration of services provided to mobile users. Through a software prototype implementation, we deploy a disaggregated multi-RAT network with services being directly accessible from the clients through the DUs of the network.

The rest of this chapter is organized as follows: Section 5.2 presents our motivation and prior proposed deployments for MEC services, and indicative related literature. In Section 5.3 we briefly present our framework for disaggregated Multi-RAT base stations and how the services are handled when placed at the fronthaul. In Section 5.4, we present evaluation metrics for the proposed solution and compare it with existing MEC placements. Finally, in Section 5.5 we conclude and present our future directions.

5.2 Motivation and Related Work

The integration and placement of computing resources as close as possible to the radio access network has attracted a lot of attention lately, with interfaces being developed specifically for integrating MEC functionality in the 5G context [40], [60]. Different methods for deploying and placing MEC-assisted services are suggested by ETSI in [78] and [48], providing guidelines for the maximum delay of the UE to service path for some state-of-the-art 5G applications (e.g. Industry 4.0, eHealth, AR/VR, etc.). Existing proposed placements are summarized in:

1. The *bump-in-the-wire* method, where MEC services are placed on the base station backhaul (link with Core Network), intercepting only data plane traffic of the cellular network and redirecting it to MEC applications.

2. Collocating MEC servers and the Core Network at the network edge. In such a case, IP traffic is intercepted beyond the Core Network and redirected to the MEC applications.

3. Using a Core Network with a *Local Breakout (LBO)* mode: control plane traffic is redirected to another Core Network instance than the data plane. MEC applications are introduced closer to the edge, collocated with the edge core network instance, and handle only data plane traffic.

A graphic representation of all the suggested placements by ETSI is illustrated in Figure 5.1. Also in this Figure, our proposed placement on the frontahul of dissagregated networks is presented.

With the emergence of Cloud-RAN, the mobile networking stack has been redefined and split at different levels [19]. Yet, the split at the higher layer 2 of the stack (between PDCP and RLC) has been standardized in the 5G NR specifications [1]. As already mentioned, these specifications define the Centralized Unit (CU) that runs the upper OSI layer 2 functions and can be instantiated at an Edge Datacenter managing one or multiple Distributed Units (DUs), that may be of heterogeneous type and integrate the lower Layer 2 and 1 functions. The disaggregation of the base stations at such layer (PDCP) allows multiple technologies to be integrated in the cell; the LTE WiFi Aggregation Adaptation Protocol (LWAAP) [4] considers the control of WiFi cells from the PDCP layer of the base station. This functionality is also drafted in the 5G-NR specifications, providing hooks for the integration of non-3GPP and legacy technologies (such as LTE) as DUs to the disaggregated cell [1]. Despite the base station disaggregation, edge deployments do not move the services closer to the DU; in the best case, the services are collocated with the CU [48], intercepting backhaul traffic (*bump-in-the-wire* method).

In this chapter, we build upon this idea, and experiment with services placed closer to the true edge, directly reachable from the DUs of the network. We leverage the work in [71] that introduces multi-technology base stations and in [74] that introduces services on the fronthaul of disaggregated LTE networks.

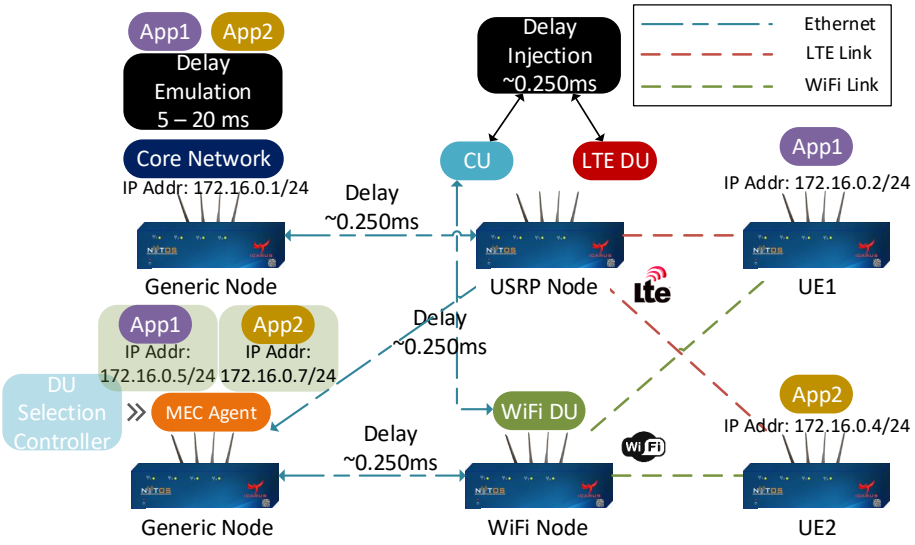


FIGURE 5.2: Overall Architecture and Experiment setup over the NITOS testbed

5.3 System Architecture

The system architecture has been designed around three entities: 1) a CU that manages multiple heterogeneous DUs, 2) DUs, orchestrating the communication with the CU and the MEC services, supporting multiple wireless access technologies, and 3) a MEC Agent which communicates with the CU for the exchange of control information and the DUs for receiving and transmitting data to the wireless network. The MEC agent can be collocated with the DUs of the network, or be deployed at another datacenter in close proximity to the multiple DUs in the network. Figure 5.2 presents our high-level architecture with the proposed placements for the MEC services. Our contributions are developed around the OpenAirInterface platform [82], which provides a software implementation of the LTE stack.

5.3.1 CU-DU design principles

According to the 5G-NR standards, disaggregated 5G base stations might address several different technologies. Hence, the standardized split is in the higher OSI stack layer 2, between PDCP and RLC layers. One CU may concurrently manage multiple DUs, but each DU is only managed from a single CU. Therefore, in such setups serving concurrently a single UE through multiple technologies can be supported, just by redirecting traffic to the DU that the CU selects. Our implementation is based on the work [71], where the authors proposed a disaggregated base station implemented in the OpenAirInterface platform. The authors introduced the F1 over IP (F1oIP) protocol for the communication between the CU and DUs. The software is managing the Service Access Points (SAP) between the mobile stack layers (*pdcp_rlc_data_request* for the Downlink (DL) traffic, and the *rlc_pdcp_data_indication* for the Uplink (UL) case). The implementation supports integration of non-3GPP DUs (e.g. a WiFi DU), by appropriate handling of the transmitted information to/from the CU.

5.3.2 DU-MEC communication

In order to incorporate services over the fronthaul, we need the appropriate interfaces between the DUs and the MEC platform that is hosting services/applications. In [74], we developed a protocol for DU to MEC communication and introduced a *MEC Agent* component. This agent generates and exchanges the appropriate messages destined to the DUs of the network, receives and delivers the respective payload destined for services hosted at the MEC site. The solution is similar to the bump-in-the-wire method proposed by ETSI, though we progress beyond that work and place the service between the CU and DUs. The MEC service can also select the technology through which each UE will be served in a per-packet basis, enabling the dynamic selection of the links that will serve each UE from the MEC's perspective. More details on this process are given in the next subsection.

5.3.3 Addressing clients over multiple technologies

As our setup is considering base stations disaggregated inside Layer 2, a mechanism is needed in order to properly address the UEs through the different DUs inside this layer. Each UE is establishing end-to-end Layer 3 connections with the Core Network or the services placed at the MEC. When the cellular UEs are attached to the network, they are addressed by the base stations using a Radio Network Temporary Identifier (RNTI). RNTIs are used to map the data plane traffic for each client to

logical, transport and physical channels. Contrary to this, non-3GPP devices (e.g. WiFi) are identified using MAC addresses. This allows them to be addressed and request services from the MEC agent based merely on the IP configuration of both sides (service and UE).

In order to cope with this incompatibility, we developed control-plane signaling as follows: whenever a new client associates with the cellular DU and a RNTI is allocated, it is broadcasted to all the different DUs and MEC agents in the system as a *rnti_inform* message. The message includes information about the RNTI of the UE, an identifier for which client of the multi-RAT network is, and through which DUs it can be served. This information is used for mapping the RNTI and the IP address allocated by the Core Network to the UE, in order to distinguish between them during the operation of the network. Similar messages are spawned whenever a UE is using a non-3GPP DU for the first time. With this approach, we ensure that all the entities of the network (MEC Agents, CU, DUs) are aware of all the clients and the DUs through which they can be served. In case of a UE using more than one technologies, the information is passed on the message, creating a mapping between the non-3GPP MAC address and RNTI used for cellular connection. Through this functionality, we can address through the MEC part of the network all the UEs that are served through the disaggregated base station architecture. This means that the MEC Agent might send the service traffic only to a set of DUs that we select, used for forwarding the traffic to the end-user.

5.3.4 MEC Service Virtualization

As mentioned, the MEC Agent orchestrates the communication of the services running on top with the DUs of the network. Whenever the agent receives MEC destined traffic, it decapsulates and injects it to the MEC service. The hosted MEC services are containerized, using Linux Containers (LXC) or Dockers, as they can be dynamically instantiated, whenever an end-user requests different services from the MEC platform. Using LXC containers has multiple benefits as it allows each hosted service to be addressed with a new address at a new container that can be migrated to other hosts if needed. As LXC places all the hosted containers under a single bridge on the edge host, the MEC agent injects traffic to this bridge, destined to the MAC address of the container implementing the requested service. Through the RNTI-IP mapping previously described, multiple UEs can use the same service, even when they are connected through different access technologies.

5.4 System Evaluation

In this section, we describe the experimental setup and findings from evaluating the proposed scheme. As we mentioned, the framework has been developed around the OpenAirInterface platform that provides a software implementation of the cellular stack. Since currently NR support in the platform is very limited, we execute our experiments using a disaggregated cell with the LTE technology. Moreover, we incorporate a WiFi device as a DU, using the software module developed in [71]. We deploy the framework at the NITOS testbed, a remotely accessible facility located in Univ. of Thessaly, Greece [72]. We employ seven nodes from the testbed as follows: 1) one equipped with a compatible SDR front-end (USRP B210) for running the LTE DU software, 2) one with a WiFi card for running the WiFi DU, 3) three generic nodes

for running the CU, OpenAirInterface Core Network and MEC Agent software respectively and 4) two more equipped with LTE dongles and WiFi cards for using them as our multi-homed UEs. Since the testbed offers an RF isolated environment, we are getting performance metrics from the deployed scheme under a controlled wireless environment.

The latency times between the nodes using the Ethernet links are approx. $250ms$ whereas we configure the wireless parameters of the channels to settings that offer to us a more stable setup (5MHz SISO mode for LTE, 40MHz IEEE 802.11n in 2.4GHz for WiFi). We experiment with two different placements of the service: 1) placement with the DUs of the network and 2) collocation with the Core Network, one of the proposed deployments by ETSI.

5.4.1 Throughput and Latency measurements

We start the evaluation of our proposed scheme with measuring the throughput that a UE obtains for reaching the services at either the fronthaul or the Core Network (Evolved Packet Core - EPC). The results are presented as a reference in Table 5.1.

TABLE 5.1: Service to UE maximum TCP throughput

Measured Path	LTE	WiFi
EPC to UE	14.19	16.76
MEC to UE	15.35	21.28

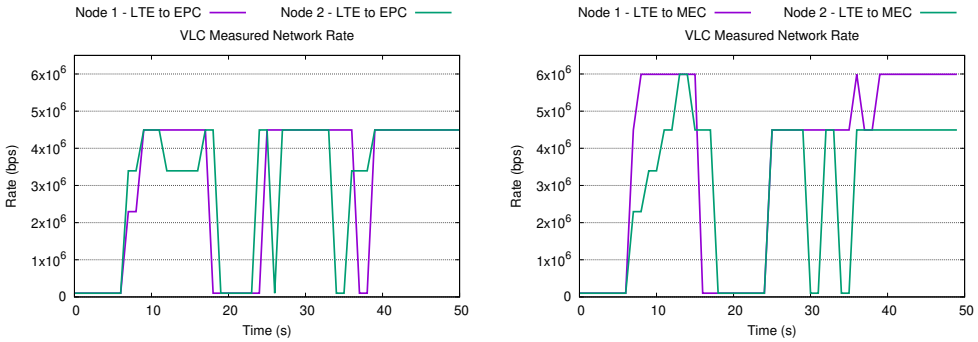
We evaluate the overall scheme for the latency time for reaching the MEC service versus a traditional deployment of the service on the EPC (or beyond), and in 5.4.2 the video streaming quality for multiple users for Dynamic Adaptive Streaming over HTTP (DASH), when placing the video server either with the DUs or at the EPC, accessed through different wireless technologies. For all our cases, we use two multihomed UEs connected to two DUs (one LTE and one WiFi) and measure on the path between the UE and the service.

TABLE 5.2: RTT Results (msec) for LTE and WiFi access to the service (Fronthaul or EPC)

	LTE to FH	WiFi to FH	LTE to EPC	WiFi to EPC
Avg. RTT	19.7	4.78	32.32	5.26
Min. RTT	15.1	4.39	26	4.59
Max. RTT	24.7	5.12	43.4	6.64

Table 5.2 shows measured averaged RTT times over the different links when placing the service at the DUs or the Core Network. Usually, in production deployments the Core Network is not placed so close to the edge as in our testbed experiments, but is preferred to be instantiated in the datacenter of the network provider. Therefore, we provide measurements for the link with/without varying latency in the backhaul (link between the CU and the Core Network). Assuming that latency is almost half of the RTT time, we see that for the cases of MEC access over LTE or WiFi, the latency is consistently less than 10ms, thus allowing several 5G applications to run, according to [60]. Moreover, we observe that WiFi outperforms LTE latency times but this is due to the less complicated design and processes that the protocol

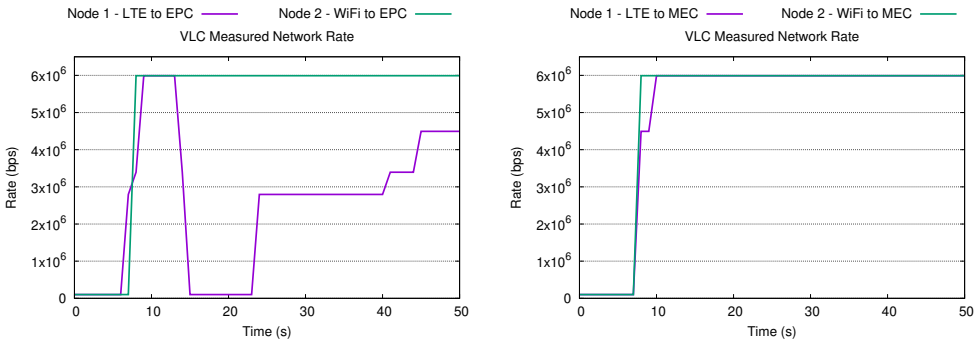
implements. Nevertheless, in a non-ideal environment other than our testbed environment, with high external interference, WiFi is prone to lower performance and thus these times may change in a non-deterministic manner.



(A) VLC Rate - 2 LTE UEs requesting at the EPC

(B) VLC Rate - 2 LTE UEs requesting at the MEC

FIGURE 5.3: VLC rates for the same access technology and service placement



(A) VLC Rate - 1 LTE and 1 WiFi UEs requesting at the EPC

(B) VLC Rate - 1 LTE and 1 WiFi UEs requesting at the MEC

FIGURE 5.4: VLC rates for different access technologies and same service placement

5.4.2 Video measurements for different placement of the service

We continue the evaluation of our scheme by running an application on different places of the network and measuring its performance. We test the network with two UEs, connected through either LTE or WiFi and request the video from a server located at the EPC or the MEC agent. For testing video services, we use a MPEG-DASH server [106], streaming videos of varying resolution (video is broken down to 1 sec segments with qualities up to 1080p). This means that for each second of the video, the client requests a segment from a set of different transcodings. The server is running through an Apache2 service, in the MEC Agent containers and the Core Network for comparing their performance. Each DASH client initially requests a Media Presentation Description (MPD) file from the server. According to the descriptions of the available video segments and the video requesting algorithm running on the application, the video is downloaded to the client. We use VLC as the end-user application, based on the policies that are described in [45]. The policy that we use for streaming the video is the following: for each video segment, VLC

estimates the channel's download rate. For the next segment to be downloaded, it requests the video with rate equal to the download rate. In the case that it does not exist (video coding rate can be lower than the channel rate), it requests the next lower representation available. For the cases that the video buffer is less than 30% occupied, the client requests the lowest available representation. Using this policy we observe the quality that each client requests, based on their view of the network.

We plot the requested video rate of the application based on its assumption of the underlying wireless channel, using the Dynamic adaptive Streaming over HTTP (DASH) capabilities of VLC player. When requesting the video from the MEC server over LTE (Fig. 5.3b), one of the two UEs manages to get video rate coded at 6Mbps, whereas the second UE is limited at maximum 4.5Mbps, which is the same for the EPC case (Fig. 5.3a). When we use different technologies (one user to LTE, one to WiFi) to request data from the EPC server (Fig. 5.4a), both clients get video coded at 6Mbps, until the LTE UE's buffer is emptied. Then it gradually starts getting better video segments up to 4.5 Mbps. From the other side, the WiFi client quickly converges to getting the best video quality available. For the case of using the same setup to get video from the MEC service, we see that both clients quickly converge to receiving the best available video quality (Fig. 5.4b).

From these results we conclude that the technology used to request the video plays a key role in the overall experience of the user. Moreover, the services that are placed on the MEC agent and therefore are closer to the UE outperform the cases of remote testbed placement (Fig. 5.3a and 5.4a).

Our experiments indicate that the access technology combined with the service placement plays a significant role in the UE application performance, thus providing insights on how the application provider could discriminate between different subscription plans of clients, allowing certain subscribers to access the same service located at the edge datacenter contrary to the rest.

5.5 Discussion and Future Work

In this chapter, we proposed, developed and evaluated a scheme for placing services over the fronthaul interface of heterogeneous 5G base stations. This placement has several benefits as it provides ground for further reduction of the UE to service latency time which enables the support for 5G applications even with legacy protocols such as LTE. Starting from a disaggregated Cloud-RAN deployment, we experimented with a novel placement of the services, for which our experimental results show up to 60% minimization in the service access latency time. Moreover, we experimented with different placements of the service and demonstrated how in conjunction with the wireless technology selection, it can impact the overall performance that a UE experiences. Given the insights from the evaluation of the porposed scheme, employing a differentiation scheme per each multi-homed UE, we are able to select the technology through which each UE will be served even per-packet basis, towards further decreasing the service latency.

An important factor that have to be taken into consideration during the placement of the MEC services or selecting the access technology, is the conditions of the wireless medium especially when the RAT operates in unlicensed bands. If the WiFi DU operates in an ultra-dense setup with other WiFi APs in the same band/channel, may create additional delays to the service access time. For this reason, in the future, we forsee extending our scheme and adding a machine learning approach on deciding dynamically which services shall be migrated to the MEC server. Through

monitoring the wireless conditions at each DU of the system and the channel quality reports from the UEs, we plan to develop a scheme in order to select and dynamically change the placement of the MEC services and the RAT through which each UE will be served, while meeting the application's requirements and ensuring the UE's QoE.

Chapter 6

Pricing in MEC Resource Allocation for 5G HetNets

Contents

6.1 Chapter Introduction

5G brings several advancements in both the air interface, the integration of legacy technologies for the formation of Heterogeneous Networks, and the utilization of edge resources. Applications developed around this ecosystem are expected to take advantage of high throughput and low latency wireless links, for supporting services around a wide domain of verticals (e.g. eHealth, Industry 4.0, AR/VR, etc.). Nevertheless, advancements in the air interface focus on enhancing the capacity of the network; low latency access is expected to be achieved through the wide application and utilization of edge computing, with resources being migrated to the network edge. In this context, and towards addressing the heterogeneity in the network access domain, ETSI revised the annotation for Mobile Edge Computing towards Multi-access Edge Computing (MEC). Through MEC, UEs in the network may use any of their available wireless network interfaces for accessing services located at the edge of the network.

At the same time, 5G introduces new concepts in the network architecture and the organization of base stations. Through the wide application of the Cloud-RAN concept, parts of the base station can be instantiated as Virtual Network Functions (VNFs) at an edge located datacenter, managing lower complexity units used for transmitting the information in the cell. This feature allows the instantiation of new base stations within an area based on demand, and is an enabler for adding heterogeneous technologies at the user access level, as a means of aggregating different

access technologies. As a matter of fact, in the recent specifications for the 5G New Radio (NR) interface, the base stations are disaggregated between the Packet Data Convergence Protocol (PDCP) and the Radio Link Control (RLC) layers, forming a Central Unit (CU) that can be instantiated in the cloud, controlling a Distributed Unit (DU) for forming the wireless cell. One CU may control multiple even heterogeneous DUs, allowing the integration of several technologies to the operator's provided cell, e.g. 5G-NR and LTE or non-3GPP based e.g. WiFi.

Although MEC is expected to play an important role in the overall 5G network operation, the placement of MEC services seems to be inherited from the legacy generations of mobile communications. In [48], ETSI provides information for all possible deployments of MEC services in the network. Nevertheless, even for the disaggregated RAN case, MEC services will be co-located with the CU at an Edge Datacenter. In [74], we provided a first experimental prototype that goes beyond these deployments, and places the provided services on the fronthaul interface of heterogeneous Cloud-RAN infrastructures, introduced in [71]. In such setups, multi-homed network users get access over multiple wireless links to services located just after the DU component of the cellular network, illustrating reduced network latency compared to conventional MEC deployments. Nevertheless, the technology through which each user may be served plays an important role in the overall perceived latency and Quality of Experience (QoE) of the mobile terminal user. Moreover, the locations for placement of the hosted services and wireless technologies used to forward data to the end-users can be exploited as differentiation parameters for charging application providers for hosting their services on the MEC platform.

In this chapter, we present a system model for resource allocation in a heterogeneous Cloud-RAN, deploying the MEC functionality at two different tiers of the network: 1) on the fronthaul interface, and 2) collocated with the Core Network. We seek to answer the following key questions:

1. How should resources from the MEC enabled network be allocated to different Service Providers?

2. How should MEC providers make use of the access technologies available for forwarding MEC data to the UEs?

3. How do these choices affect the service-to-UE latency?

The rest of the chapter is organized as follows: Section 6.2 is presenting a literature overview in the field. Section 6.3 presents our system model for the resource allocation across the different MEC tiers, and Section 6.4 details our approach for access technology selection. In Section 6.5 we present the evaluation of our proposed scheme and showcase our findings, whereas in section 6.6 we conclude the chapter and present some future directions.

6.2 Related Work

As the application of MEC is designed to deliver low latency for service access, it has received a great level of attention in the recent specifications for 5G and in relevant research. In [48], ETSI specifies the different deployments for MEC services starting from the 4G network architecture and its evolution for 5G. This white paper summarizes the different interfaces needed for hosting services over a MEC enabled server, and specifies the deployments as follows: 1) the *bump-in-the-wire* mode, where the

service is located just after the base station, intercepting data-plane traffic, and relieving the network from the extra delay added for sending traffic to the Core Network, 2) the case of collocating the MEC services with the Core Network at an Edge datacenter, which has the benefit of handling IP traffic just after the Core Network, and 3) the *local break-out mode* where a part of the Core Network is handling only data plane traffic collocated with the base station, whereas the control plane traffic is sent to a traditional Core Network deployment. The advantage of the third solution is that it blends the benefits of the two prior solutions, but requires increased complexity on the core network implementation. In [60], ETSI specifies all the different interfaces enabling the MEC operation for different components of the network.

Based on the disaggregated model of a base station, according to the Cloud-RAN concept, in [74] we introduced a new deployment for the MEC services; since the base station is disaggregated in two components, we developed and evaluated a prototype illustrating the traffic flow as UE-DU-MEC, instead of UE-DU-CU-MEC that ETSI specifies as the *bump-in-the-wire* method for Cloud-RAN. The implementation is based on the Open Source OpenAirInterface platform [82], and extends our prior contributions for integrating non-3GPP technologies in the cell [71]. Thus, the solution provides a first effort for enabling a MEC platform, with the services being deployed as close as possible to the network edge. The prototype showcased low latency for MEC service access, able to achieve less than 10ms for a standard LTE cell in the access network. This solution is adopted in this work as well, as the base of our experimental platform.

Similar solutions for deploying the services on the edge exist in such experimental platforms. For example, in [52], the authors use the OpenAirInterface platform in order to deploy services co-located with the Core Network. By using an SDN approach just after the Core Network, the authors provided low-latency times for accessing hosted services for specific UEs. Similarly, in [66], the authors implement the *bump-in-the-wire* method on the same platform. This prototype may achieve low latency times, but as it is solely implemented in application space, it strives to provide real time services for high-load cells. In [108], the authors present all the possible enablers for MEC operation when multiple technologies are used for user access. When considering the existence of multiple paths in the wireless part of the network, allocating the network resources needs to be revisited. For example, in [35], the authors deal with the Radio Access Technology (RAT) association problem for Heterogeneous Networks (HetNets) when MEC resources are present. Similarly, in [116] the authors consider a multi-RAT network with MEC resources, and attempt to minimize the overall energy consumption of the network with a hollistic approch. Application specific MEC enhancements are also presented in [46]. The authors use dynamic adaptive streaming video over a MEC service, extended to ensure the optimal QoE for the end-users.

In this work, we initially model a MEC platform in terms of resource allocation. We use a pricing scheme to determine how the resources residing at the MEC platform (CPU, memory and storage) shall be allocated to Service Providers (SPs). Subsequently, we introduce an algorithm for selecting the network access technology used to serve each user of the network, in order to ensure that the overall service access latency times are kept low. Finally, we employ testbed experimentation with the objective to evaluate our framework for different placements of the MEC service: 1) on the fronthaul interface of the multi-technology Cloud-RAN, 2) on the Core Network, and 3) deployed at a remote datacenter.

6.3 MEC Pricing Scheme

We consider a two-stage MEC pricing scheme, where the MEC owner (operator) decides the price p per unit of MEC bundled resources (CPU, memory and storage) in the first step and in the second step the service/content providers, interested in providing low latency services, decide the level of bundled MEC resources, which they intend to pay as a function of the price and the latency sensitivity of the provided service. We approach the pricing problem using backward induction following the rationale of [75], examining first the service/content providers' demands (Stage II) and then the MEC operator's decision on the price (Stage I). We propose two pricing models, one linear in Section 6.3.1 and one exponential in Section 6.3.2.

6.3.1 MEC Resource Allocation With Linear Pricing

Stage II: The payoff function of the Service Provider SP_i, $i = 1, ..., N$, for acquiring b_i units of MEC bundled resources with a price p per bundle unit, following the linear pricing model, is expressed as

$$U_i^{lin}(b_i) = ln(1 + \theta_i b_i) - p b_i \tag{6.1}$$

with θ_i representing the normalized latency sensitivity of SP_i, $\theta_i \in [0, 1]$. This payoff function of SP_i is equal to the logarithmic utility function, that expresses the diminishing return of getting additional resources, minus the linear price that SP_i has to pay for acquiring b_i quantity of MEC resources. We notice that $U_i^{lin}(b_i)$ is a concave function, since $U(b_i)'' = -(\theta_i/(1 + \theta_i b_i))^2 < 0$. Thus, it has only one maximum, and therefore the local maximum is also the global maximum. Differentiating (6.1) we have

$$\frac{\partial U_i^{lin}}{\partial b_i} = \frac{\theta_i}{1 + \theta_i b_i} - p = 0 \tag{6.2}$$

The optimal value of MEC resources that maximizes SP_i's payoff is

$$b_i^* = \begin{cases} \frac{1}{p} - \frac{1}{\theta_i}, & \text{if } p \leq \theta_i \\ 0, & \text{otherwise} \end{cases} \tag{6.3}$$

Stage I: We assume that the N SPs that are requesting for MEC resources present similar latency sensitivity. Otherwise, no need to purchase MEC resources would exist. Thus, we assume that their latency requirements are such that $\max(\theta_i) - \min(\theta_i) < \varepsilon$, where $\varepsilon > 0$. Under this assumption, the MEC owner's choice of price p is such, that the SP with the $\max(\theta_i)$ is allocated the maximum value of MEC resources $b_{\max}$, aiming to provide the best available service to SPs with higher latency sensitivity compared to the rest of the SPs requesting resources from the MEC agent. We also assume that the MEC agent has adequate available resources to satisfy the requests of all SPs under consideration. The price is formed according to (6.4).

$$p = \frac{\max(\theta_i)}{1 + \max(\theta_i) b_{\max}} \tag{6.4}$$

The provider aims to give to every SP_i the opportunity to have access to the MEC resources. This means that even for the SP with the $\min(\theta_i)$, the quantity $1/p - 1/\min(\theta_i)$ is positive. Using (6.4) we find the range of values of ε under which this

MEC resource allocation is feasible. This range is expressed as

$$0 < \varepsilon \leq \max(\theta_i) \min(\theta_i) b_{\max} \tag{6.5}$$

The allocated level of MEC resources to each SP_i following the linear pricing model is expressed as

$$b_i = \frac{1 + \max(\theta_i) b_{\max}}{\max(\theta_i)} - \frac{1}{\theta_i} \tag{6.6}$$

6.3.2 MEC Resource Allocation With Exponential Pricing

For the MEC resource allocation with exponential pricing, we follow the same steps as described in the linear pricing approach.

Stage II: The payoff function of SP_i under the exponential pricing model, for acquiring b_i units of MEC bundled resources is expressed as

$$U_i^{exp}(b_i) = ln(1 + \theta_i b_i) - p_e(e^{b_i} - 1) \tag{6.7}$$

We notice that $U_i^{exp}(b_i)$ is a concave function, since $U_i^{exp}(b_i)'' = -\left(\theta_i/(1+\theta_i b_i)\right)^2 - p_e e^{b_i} < 0$. Thus, it has only one maximum, and therefore the local maximum is also the global maximum. Differentiating (6.7) we have

$$\frac{\partial U_i^{exp}}{\partial b_i} = \frac{\theta_i}{1 + \theta_i b_i} - p_e e^{b_i} = 0 \tag{6.8}$$

We express (6.8) as

$$ln\left(\frac{1}{p_e}\right) + \frac{1}{\theta_i} = \left(b_i + \frac{1}{\theta_i}\right) + ln\left(b_i + \frac{1}{\theta_i}\right) \tag{6.9}$$

For $x = b_i + \frac{1}{\theta_i}$ and $y = ln\left(\frac{1}{p_e}\right) + \frac{1}{\theta_i}$, (6.9) can be written as

$$y = x + lnx \tag{6.10}$$

which can be also expressed as

$$xe^x = e^y \tag{6.11}$$

Taking the value of the Lambert W function [26] of each part of (6.11) and using the Lambert W function identity $W(xe^x) = x$, we have $x = W(e^y)$. Replacing x and y we have

$$b_i^* = \begin{cases} W\left(\frac{e^{\frac{1}{\theta_i}}}{p_e}\right) - \frac{1}{\theta_i}, & \text{if } \theta_i \geq \frac{1}{W\left(\frac{e^{\frac{1}{\theta_i}}}{p_e}\right)} \\ 0, & \text{otherwise} \end{cases} \tag{6.12}$$

Stage I: The price p_e that the MEC owner decides in the exponential pricing model is such, that SP_i with $\max(\theta_i)$ is allocated the maximum value of MEC resources $b_{\max}$. The price is formed according to (6.13).

$$p_e = \frac{\max(\theta_i)}{(1 + \max(\theta_i) b_{\max})\, e^{b_{\max}}} \tag{6.13}$$

As the provider aims to give to all N SPs the opportunity to to have access to the MEC resources, the level of resources that will be allocated to the user with the

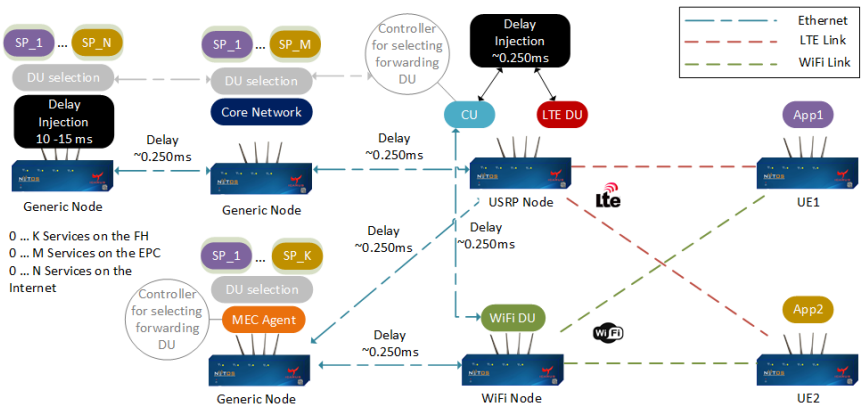


FIGURE 6.1: Experimental topology for evaluating our scheme; controllers residing at the CU and the MEC agent part select the forwarding DU(s) for serving each UE in a per-packet basis. Services are placed either on the Fronthaul (FH), Core Network (EPC) or emulated Internet.

$\min(\theta_i)$ should also be positive. This means that the range of latency sensitivity of the N SPs is such, that

$$W\left(\frac{e^{\frac{1}{\min(\theta_i)}}}{p_e}\right) - \frac{1}{\min(\theta_i)} > 0 \tag{6.14}$$

The allocated resources to each SP_i following the exponential pricing model is expressed as

$$b_i = W\left(\frac{(1+\max(\theta_i)b_{\max})e^{b_{\max}+\frac{1}{\theta_i}}}{\max(\theta_i)}\right) - \frac{1}{\theta_i} \tag{6.15}$$

6.4 System Architecture

In this section we describe the components that we are utilizing in order to port our proposed scheme in a real-world setup. In addition, the procedure for the selection of the RAT through which each UE will be served is presented in 6.4.2.

6.4.1 System Components

As our starting system architecture, we use a disaggregated multi-technology Cloud-RAN base station, extensively described in [71]. In such a setup, we distinguish the following components and roles:

1. The **CU**, which is running the higher layer 2 functions of the base station (PDCP layer and upwards), and provides the interface to the Core Network.

2. The **3GPP DU**, running the lower layer 2 functions of the base station (RLC and below), and performs the transmission of the traffic over the air. The DU may support heterogeneous technologies (e.g. 5G NR or LTE) and uses the Radio Network Temporary Identifiers (RNTI) for addressing each client.

3. The **non-3GPP DU**, which is running the layer 2 functions for non 3GPP technologies (e.g. WiFi) and communicates with the CU for sending/receiving traffic to/from the wireless network. As it communicates directly with the

PDCP layer of the CU, it handles and encapsulates the data in the appropriate format. MAC addresses are utilized for addressing each client.

4. The **Core Network**, which is the entry and exit point for user plane data to the base station network.

5. The **MEC agent**, which is enabling the data exchange from services located at the Fronthaul interface with the DUs directly, without using the CU as intermediary node.

6. The **Technology Selection** modules, which reside on the MEC agent and the CU, and are able to select the forwarding DU(s) for each client of the network.

7. The **Hosted Services** over this heterogeneous infrastructure, which are containerized services running on top of the MEC agent, the Core Network or any other remote datacenter. The container technology we use is LXC.

Figure 6.1 shows how these components have been mapped to a real testbed setup. We employ the NITOS testbed [83], which provides all the required experimental components for supporting our experimentation.

Algorithm 5 MEC side selection of RATs for each UE.

```
Calculate the resource allocation of the services
based on the chosen pricing (Linear or Exponential)
while 1 do
    for each service request do
        if Both DU meet the service's requirements then
            Choose the DU with the lowest Latency
            if DU capacity is lower than 50% then
                Use both DU with percentage q and
                (100-q) of the time respectively
            end if
        else
            Choose the DU which meets the requirements
        end if
        Send to the proper REST API:
        the UE id and DU id/s
    end for
    Calculate the resource allocation of the services
    based on the chosen pricing (Linear or Exponential)
end while
```

6.4.2 Selection of Radio Access Technology

As MEC considers multiple wireless technologies for service access, network selection is an issue of paramount importance. The last-hop link used for serving each user may exemplify different access times, based on factors such as the cell coverage, the location of the UE in the cell, the allocated modulation and coding scheme, the load of the cell and external interference, especially for non-3GPP technologies such as WiFi. In this section, we introduce an algorithm for selecting the last-hop wireless connection in a per client basis for each UE. We assume at this point that each UE is

multi-homed and is using all of its available technologies to communicate with the MEC and Core Network.

Although our scope is to minimize service latency times, we bear in mind the different capacities of the wireless networks used to serve each UE. Therefore, each UE might be concurrently served by combinations of the available technologies, while the per-packet traffic latency is kept below a threshold limit. In the case that the capacity of a technology is about to be reached, the algorithm might choose to serve an end-user by another technology. Algorithm 5 shows how the MEC part of the network makes these selections.

In order for the decisions for each forwarding DU to be applied, separate controllers have been developed at two different points: 1) on the MEC agent, which is handling the MEC traffic on the fronthaul interface, and 2) on the CU side of the network, that handles the traffic before being sent to each DU. Both the controllers operate under the same principle. They expose a REST API that gets as inputs the identifier for each UE of the network and the DU or combination of DUs that will be used for forwarding the traffic. For the case of combination of DUs, defining the percentage of traffic for each DU is also supported. Based on this, our algorithm operates as follows. We initially calculate the resource allocation for the service providers based on the pricing model. For each new UE service request, the controllers residing at the MEC agent or the CU select the forwarding DU that meets the specific UE requirements. If all DUs are able to serve this UE then the DU with the lower value of latency is selected. If the capacity of the DU with lower access latency is not exceeding 50%, the service request is served through this DU. In the case that this threshold is exceeded, we split the traffic over multiple DUs with 50% transferred over the WiFi DU and 50% over the LTE. The information is subsequently sent to the respective controllers, managing the forwarding of the data to the UEs.

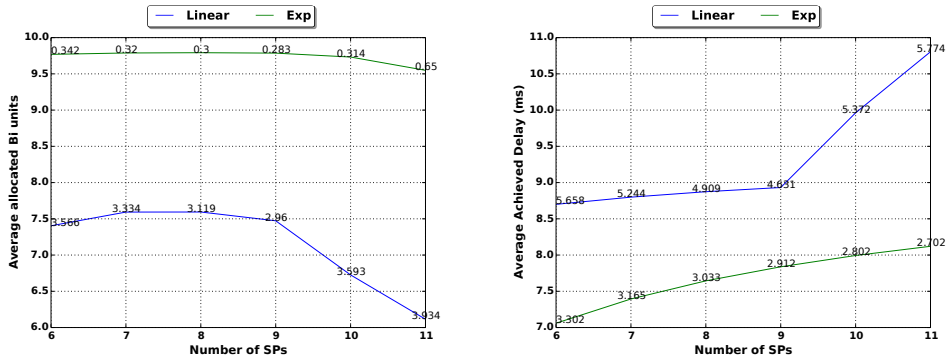
6.5 System Evaluation

In this section, we present our experimental findings. First, we present the system components and selected configurations and then we showcase results of the experiments. We employ the NITOS testbed for extracting the latency values for the service placement for two MEC setups (service on the Fronthaul-FH or on the Core Network-EPC) and the Internet.

TABLE 6.1: System Latency times for the different placements of the MEC service and the RAT

Service Location	**WiFi**	**LTE**
Fronthaul (FH)	2.39	9.85
Core Network (EPC)	2.63	16.16
emulated Internet	12.57	25.9

We employ two different wireless technologies for accessing the MEC services, either LTE for 3GPP access or WiFi for non-3GPP access, both with the same configuration: 2x2 MIMO and 20MHz channel bandwidth. Our testbed setup can accommodate multiple DUs, including 5G-NR, but as the OAI 5G-NR platform is currently in development state we omit this wireless technology from our evaluation. The overall setup is shown in Figure 6.1 and Table 6.1 presents the achieved access latency for



(A) Average assigned b_i units per SP (B) Average achieved delay per SP

FIGURE 6.2: Experimental results for the 1st scenario of Service Providers allocated to the system (low demand)

the different service placements for all the available RATs. We also consider 6 different types of services based on their requirements in terms of latency and throughput derived from [95].

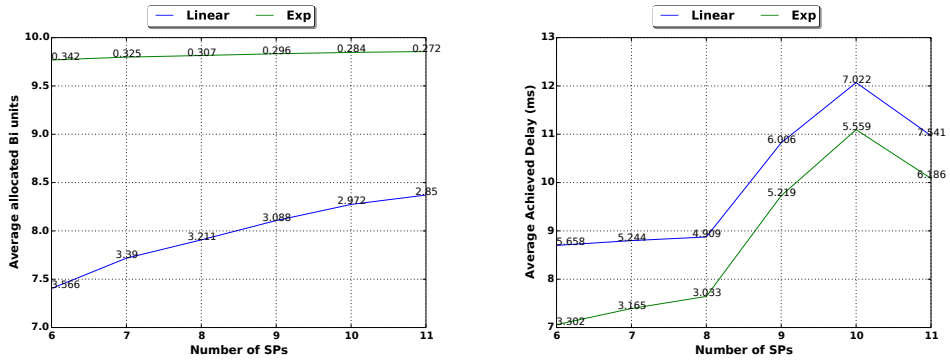
We create a matching between the types of service and θ_i values, as shown in Table 6.2 and for each service provider, we select a value uniformly from the provided ranges per each SP. Table 6.3 shows the MEC placements and RAT allocation for each of the different services that we use for both pricing models. As we can see, the exponential pricing is more flexible and can serve all the services through MEC, whereas the linear pricing cuts-off the service of lower latency sensitivity from the MEC. This occurs due to the fact that in the linear model, the range of values that renders the MEC resource allocation feasible in (6.5) is more strict than the respective (6.14) of the exponential model, for the same $b_{\max}$ and θ_i.

TABLE 6.2: Applications Normalized Latency Sensitivity

APPLICATION TYPE	**Initial Selected θ_i**	**θ_i range**
AR/VR	0.95	[0.85 - 1]
V2X	0.8	[0.7 - 0.85)
VIDEO STREAM	0.65	[0.55 - 0.7)
VoIP	0.5	[0.3 - 0.55)
BROWSING	0.2	[0.1 - 0.3)
MAIL	0.05	[0.0 - 0.1)

We evaluate our proposed scheme with two different use cases. Each use case is examined for the two provided pricing models under the same scenario. We measure the allocated resource bundles for each SP, for each new SP that enters the system, and the aggregate latency for the network UEs accessing the provided services. We present the average performance of our performed experiments repeated 100 times, along with the standard deviation for each measurement. The two use cases are differentiated regarding the level of resources that each new SP demands. In the first use case, services with low demands are introduced such as VoIP, web services or e-mail servers, whereas in the second, services with high demands are introduced, such as AR/VR, V2X and video streaming. For both cases, we plot the resource allocation and delay after the initial placement of six different SPs, following the application types of Table 6.2.

Figure 6.2a shows the average allocation of bundled MEC resource units b_i for each pricing model as the number of SPs increases along with the standard deviation. We observe that with linear pricing, the allocation of b_i units depends highly



(A) Average assigned b_i units per SP (B) Average achieved delay per SP

FIGURE 6.3: Experimental results for the 2nd scenario of Service Providers allocated to the system (high demand)

on the SPs' requirements, in contrast to the exponential pricing where all SPs are assigned with almost equal and higher level resource bundles. This happens mainly because with exponential pricing, MEC resources are more evenly spread over the SPs requesting to place their services on the MEC (FH or EPC). In Figure 6.2b, we observe a similar trend for the two pricing models for the average achieved delay of the system. The exponential pricing achieves lower average latency, as it more evenly

TABLE 6.3: Service Providers allocations for the different pricing models

SP ID	Linear	Exp	RAT
AR/VR	FH	FH	WiFi
V2X	FH	FH	WiFi
VIDEO STREAM	EPC	EPC	Both
VoIP	EPC	EPC	Both
BROWSING	EPC	EPC	Both
MAIL	Internet	EPC	Both

places services to the MEC, presenting lower levels of average delay. In the second use case, the average latency per each SP is lower than the linear policy. Comparing the delay of the linear pricing policy with the first use case experiment, we observe higher latency times. This is happening because the types of SPs used for this experiment pose higher demands regarding their latency requirements. Due to this fact, the assignment of multiple RATs per service is employed only for the SPs with low demands, contrary to the first use case where most of the services use multiple RATs. The rest of the services are assigned to one RAT only (LTE), once the capacity of the RAT with the minimum delay (WiFi) is fully allocated. Based on the findings in Table 6.1, the difference in latency between the WiFi and LTE is high; this is also reflected in the average achieved delay of the system in Figure 6.3b.

6.6 Chapter Conclusion and Future Work

In this chapter, we presented a scheme for the resource allocation of MEC resources to different service providers using two pricing policies. We modeled our approach and used a testbed setup to evaluate our scheme, using two different placements of the MEC services: on the fronthaul interface of Cloud-RAN base stations, or collocated with the Core Network. Our proposed scheme employ two heterogeneous

Distributed Units (LTE, WiFi) for providing access to the end-users. Through the integration of multiple technologies at the base station level, we are able to achieve differentiation for the latency access times for each service per each network UE. Our experiments denote that through our approach, MEC resources can be allocated while the average latency per each SP can be kept below a threshold, by utilizing multiple links at the same time.

In the future, we foresee extending our scheme towards modeling the access of each UE in the network from the SP's perspective, even for the cases of UEs accessing the same service but under different agreements with the operator. Moreover, we plan to integrate migration of the services in the system, based on the mobility patterns detected for each UE. This follow-me approach will have to take into consideration the wireless conditions on the invovled DUs and the perceived channel qualities from the UE side in order also to decide whether or not it will have to change the placement (fronthaul, backhaul) of the MEC service during the migration.

Chapter 7

Conclusions and Future Work

Contents

7.1 Summary of the Contributions

In this thesis, we focused on the resource allocation problem in Heterogeneous Networks and how each multi-homed client should select the RAT through which it will be served. Moreover, we studied the placement of MEC services on the fornthaul of a Cloud-RAN, utilizing multiple RATs for serving the end-users. We further examined different pricing models for the MEC resources in such networks. These studies gave insights for the infrastructure providers on how they could offer different bundles of their resources to content providers depending on their services' requirements. All of our contributions were not only based on mathematical formulation and analysis relying on the optimization theory, but they were strengthened by real system implementations, which enable us to evaluate our solutions under realistic environment settings.

7.1.1 Pricing scheme in 5G HetNets

The next topic that we focused on was the heterogeneous ultra-dense networks and how different resource allocation algorithms could be applied in such environments. We build upon the Paris Metro Pricing and design a scheme with dynamic prices, as a policy for selecting a RAT when operating inside a multi-RAT environment. We formulated the problem and defined the utility functions of each client for accessing the target RAT and we investigated its performance for several mobility scenarios. For the evaluation of our design both simulations and real testbed experiments were used, investigating differnet mobility patterns for the users.

7.1.2 Resource allocation for Multi-homed Clients and Multiple Traffic Classes in 5G HetNets

Following this, we extended this resource allocation algorithm by enabling an end-user to be concurrently connected to multiple RATs and to assign each one of its applications to different RAT towards maximizing its utility function. Through the validation of our proposed scheme, we incurred the importance of the UE's ordering during the update period (system cycle) where the UEs receive the new system status and decide whether or not will change its traffic classes allocation. Subsequently, we examined three different ordering policies (based on data rate demands, spectrum efficiency, and sensitivity) and presented experimental results obtained from the application of our proposed model in a real testbed environment. The results showed how the ordering policies may affect the cost incurred at each UE as well as the utilization of the available RATs.

7.1.3 Dynamic Resource Allocation scheme in 5G HetNets

Finally, based on the findings from our resource allocation algorithm in a 5G HetNet when differnet UE ordering policies are applied, we studied their fairness compared to a weighted proportionally fair bandwidth allocation policy (reference policy). Given the obtained results of this evaluation, we proposed a new policy selection scheme that in each system cycle, the Network Controller dynamically select the appropriate policy from the predefined policy set (based on data rate demands, spectrum efficiency, and sensitivity). In each cycle, the selected policy presents the closest possible performance to the reference policy, identified through its Kendall tau correlation coefficient. The initial pricing was based on the availability of access technologies' bandwidth but given the obtained results from the evaluation it was further extended to include the rate demands of the traffic classes of UEs. Both simulations and testbed experimentation were employed for the evaluation of the proposed framework and enabled the assessment of the bandwidth allocation efficiency, the variation of access prices, the policy fairness and the induced cost to UEs covered by the HetNet system.

7.1.4 MEC placement in Heterogeneous Cloud-RANs

Initially, we studied the placement of services at the edge of a Heterogeneous Cloud-RAN. We experimented with different placements of the MEC service, either on the fronthaul (collocated with the Cloud-RAN Base Station) or the backhaul (collocated with the Core Network) and for different access technologies (LTE, WiFi). Our goal was to minimize the service access latency for the end-users something that was verified through the evaluation of our solution. Moreover, the experimental results gave insights on how the UE's QoE is affected from the service placement in conjuction with the wireless technology that it will be served.

7.1.5 MEC pricing in 5G HetNets

Subsequently, we designed a scheme for the resource allocation of MEC resources to different service providers using two pricing policies. In order to evaluate our proposed scheme under a real-world environment, we utilized our MEC implementation in the NITOS testbed, using two different placements of the services: on the fronthaul interface of Cloud-RAN base stations, or collocated with the Core Network. Knowing from our studies that different RATs don't achieve the same latency

times, enable us to propose different pricing for each service per each end-user. Our experiments showed that utilizing multiple RATs at the same time and with the appropriate placement of the MEC service, the average latency can be kept below a threshold and meet the applications' requirements.

7.2 Future Work

In the future, we foresee extending our resource allocation scheme in 5G HetNets by employing artificial intelligence classification algorithms that will enable UEs to further automate the access technology selection decisions aiming to minimize the induced access costs and evaluate it under novel mobility schemes for the clients entering and leaving the Multi-RAT system. Another direction that we plan to investigate invovles both the resource allocation scheme and the MEC service placement. We plan to study the modeling of the access of each UE in the network from the Service Provider's perspective, even for the cases of UEs accessing the same MEC service but under different agreements with the operator. Moreover, we plan to include a machine-learning process for the live migration of the MEC services to different edge hosts. The decision for the migration will be taken based on the traffic patterns of each UE that will be monitored at the Core Network, and taking into consideration the wireless conditions at each DU and reported channel qualities from the clients. Finally, we aim to extend the MEC resource pricing scheme by including the energy cost for each bundle of MEC resources. In addition, we would like to expand the profiles of the supported applications in the MEC host as resource bundles, depending on the hosted service and apply further pricing policies which will take into consideration other requirements as metrics except from latency solely.

Bibliography

[1] 3GPP. *"3GPP TS 38.473 V15.1.1 (2018-04), 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NG-RAN; F1 application protocol (F1AP) (Release 15)"*. 2017.

[2] 3GPP. *3GPP TR 22.864 V15.0.0 (2016-09), 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers - Network Operation; Stage 1 (Release 15)*. 2016.

[3] 3GPP. *3GPP TR 38.806 V15.0.0 (2017-12), 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study of separation of NR Control Plane (CP) and User Plane (UP) for split option 2; (Release 15)*. 2017.

[4] 3GPP. *3GPP TS 36.360 V14.0.0 (2017-03), 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE-WLAN Aggregation Adaptation Protocol (LWAAP) specification (Release 14)*. 2017.

[5] 3GPP. *3GPP TS 38.401 V0.2.0 (2017-07), 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NG-RAN; Architecture description (Release 15)*. 2017.

[6] 3GPP. *3GPP TS 38.470 V0.3.0 (2017-09), 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NG-RAN; F1 general aspects and principles (Release 15)*. 2017.

[7] A. Aissioui et al. "On Enabling 5G Automotive Systems Using Follow Me Edge-Cloud Concept". In: *IEEE Transactions on Vehicular Technology* 67.6 (June 2018), pp. 5302–5316. ISSN: 1939-9359. DOI: 10.1109/TVT.2018.2805369.

[8] E. Aryafar et al. "RAT Selection Games in HetNets". In: *Proc. IEEE INFOCOM 2013*. 2013, pp. 998–1006.

[9] Alaa Awad et al. "Network association with dynamic pricing over D2D-enabled heterogeneous networks". In: *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2017, pp. 1–6.

[10] Aruna Balasubramanian, Ratul Mahajan, and Arun Venkataramani. "Augmenting Mobile 3G Using WiFi". In: *Proc. of the 8th ACM International Conference on Mobile Systems, Applications, and Services*. June 2010.

[11] Hamzeh Beyranvand et al. "Toward 5G: FiWi Enhanced LTE-A HetNets With Reliable Low-Latency Fiber Backhaul Sharing and WiFi Offloading". In: *IEEE/ACM Trans. Netw.* 25.2 (Apr. 2017), pp. 690–707. DOI: 10.1109/TNET.2016.2599780.

[12] G. Bianchi. "Performance Analysis of the IEEE 802.11 Distributed Coordination Function". In: *IEEE JSAC* 18.3 (Mar. 2000), pp. 535–547.

[13] *Blu Wireless Technology WiGig*. URL: http://www.bluwirelesstechnology.com/wigig/.

[14] E. Borcoci et al. "Optimization of Multi-server Video Content Streaming in 5G Environment". In: *The Eighth International Conference on Evolving Internet INTERNET 2016*. 2016.

[15] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[16] L. Breslau et al. "Web Caching and Zipf-like Distributions: Evidence and Implications". In: *Proc. of IEEE INFOCOM 1999*. Mar. 1999.

[17] P. Chander and L. Leruth. "The optimal product mix for a monopolist in the presence of congestion effects: A model and some results". In: *International Journal of Industrial Organization* 7.4 (1989), pp. 437–449.

[18] Chi-Kin Chau, Qian Wang, and Dah-Ming Chiu. "On the Viability of Paris Metro Pricing for Communication and Service Networks". In: *Proc. IEEE INFOCOM 2010*, pp. 1–9.

[19] A. Checko et al. "Cloud RAN for mobile networks - a technology overview". In: *IEEE Communications surveys & tutorials* 17.1 (2015).

[20] Lin Chen. "A distributed access point selection algorithm based on no-regret learning for wireless access networks". In: *2010 IEEE 71st Vehicular Technology Conference*.

[21] Y. Chen et al. "Joint user association and resource allocation in the downlink of heterogeneous networks". In: *IEEE Transactions on Vehicular Technology* 65.7 (2016), pp. 5701–5706.

[22] M. H. Cheung and J. Huang. "Optimal Delayed Wi-Fi Offloading". In: *Proc. of 11th IEEE International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt '13)*. May 2013.

[23] I. Chih-Lin et al. "New paradigm of 5G wireless internet". In: *IEEE Journal on Selected Areas in Communications* 34.3 (2016), pp. 474–482.

[24] I. Chih-Lin et al. "NGFI, the xHaul". In: *Globecom Workshops (GC Wkshps)*. IEEE. 2015, pp. 1–6.

[25] Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020*. Feb. 2016.

[26] Robert M Corless et al. "On the Lambert W Function". In: *Advances in Computational Mathematics* 5.1 (1996), pp. 329–359.

[27] P. Coucheney, C. Touati, and B. Gaujal. "Fair and Efficient User-Network Association Algorithm for Multi-Technology Wireless Networks". In: *Proc. IEEE INFOCOM 2009*. 2009, pp. 2811–2815.

[28] S. Deb et al. "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets". In: *IEEE/ACM transactions on networking* 22.1 (2014), pp. 137–150.

[29] C. Desogus et al. "A Traffic Type-Based Differentiated Reputation Algorithm for Radio Resource Allocation During Multi-Service Content Delivery in 5G Heterogeneous Scenarios". In: *IEEE Access* 7 (2019), pp. 27720–27735.

[30] NTT DoCoMo. "5G radio access: Requirements, concept and technologies". In: *White Paper, Jul* (2014).

[31] Z. Du et al. "Exploiting User Demand Diversity in Heterogeneous Wireless Networks". In: *IEEE Transactions on Wireless Communications* 14.8 (Aug. 2015). ISSN: 1536-1276. DOI: 10.1109/TWC.2015.2417155.

[32] S. Dutta et al. "On-the-fly QoE-aware Transcoding in the Mobile Edge". In: *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE. 2016, pp. 1–6.

[33] J.P. Ebert et al. "Measurement and Simulation of the Energy Consumption of an WLAN Interface". In: *Technical University of Berlin, Telecommunication Networks Group, Tech. Rep. TKN-02-010* (June 2002).

[34] M. Emara, M. C. Filippou, and D. Sabella. "MEC-Assisted End-to-End Latency Evaluations for C-V2X Communications". In: *2018 European Conference on Networks and Communications (EuCNC)*. June 2018, pp. 1–9. DOI: 10.1109/EuCNC.2018.8442825.

[35] M. Emara, M. C Filippou, and D. Sabella. "MEC-aware cell association for 5G heterogeneous networks". In: *Wireless Communications and Networking Conference Workshops (WCNCW), 2018 IEEE*. IEEE. 2018, pp. 350–355.

[36] Ericsson. *5G Radio Access*. [Online],http://www.ericsson.com/res/docs/whitepapers/wp-5g.pdf. 2011.

[37] JJ Escudero-Garzás and Carlos Bousoño-Calzón. "An Analysis of the Network Selection Problem for Heterogeneous Environments with User-Operator Joint Satisfaction and Multi-RAT Transmission". In: *Wireless Communications and Mobile Computing* 2017 (2017).

[38] ETSI. *3GPP TS 23.261: IP flow mobility and seamless Wireless Local Area Network (WLAN) offload; Stage 2 (V11.0.0)*. Sep. 2012.

[39] ETSI. *ETSI GS MEC 009 V1.1.1 (2017-07), Mobile Edge Computing (MEC); General principles for Mobile Edge Service APIs*. 2017.

[40] ETSI. *ETSI GS MEC 011 V1.1.1 (2017-07): Mobile Edge Computing(MEC); Mobile Edge Platform Application Enablement*. 2017.

[41] ETSI. *ETSI GS NFV-MAN 001 V1.1.1 (2014-12), Network Functions Virtualisation (NFV); Management and Orchestration*. 2014.

[42] TS ETSI. "123 107 V13.0.0 (2016-01) Digital cellular telecommunication system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Quality of Service (QoS) concept and architecture (3GPP TS 23.107 version 13.0.0 Release 13)". In: *ETSI Technical Specification* (2016).

[43] R. Fagin, R. Kumar, and D. Sivakumar. "Comparing top k lists". In: *SIAM Journal on Discrete Mathematics* 17.1 (2003), pp. 134–160.

[44] S. Floyd and V. Jacobson. "Random early detection gateways for congestion avoidance". In: *IEEE/ACM Transactions on Networking (ToN)* 1.4 (1993), pp. 397–413.

[45] F. Fund et al. "Performance of DASH and WebRTC Video Services for Mobile Users". In: *2013 20th International Packet Video Workshop*. Dec. 2013, pp. 1–8. DOI: 10.1109/PV.2013.6691455.

[46] C. Ge et al. "QoE-driven DASH video caching and adaptation at 5G mobile edge". In: *Proceedings of the 3rd ACM Conference on Information-Centric Networking*. ACM. 2016, pp. 237–242.

[47] A. Ghosh et al. "Heterogeneous Cellular Networks: From Theory to Practice". In: *IEEE Communications Magazine* 50.6 (2012).

[48] F. Giust et al. *ETSI White Paper No. 24: MEC Deployments in 4G and Evolution Towards 5G*. 2018.

[49] Bo Han et al. "Cellular Traffic Offloading Through Opportunistic Communications: A Case Study". In: *Proc. of the 5th ACM Workshop on Challenged Networks*. Sept. 2010.

[50] Red Hat. *libvirt: The virtualization API*. 2012.

[51] Linhai He and Jean Walrand. "Pricing Differentiated Internet Services". In: *Proc. IEEE INFOCOM 2005*. Vol. 1, pp. 195–204.

[52] Anta Huang et al. "Low latency MEC framework for SDN-based LTE/LTE-A networks". In: *Communications (ICC), 2017 IEEE International Conference on*. IEEE. 2017, pp. 1–6.

[53] J. Huang et al. "A Close Examination of Performance and Power Characteristics of 4G LTE Networks". In: *Proc. of the 10th ACM International Conference on Mobile Systems, Applications, and Services*. June 2012.

[54] S. Ibnalfakih, E. Sabir, and M. Sadik. "Multi-homing as an Enabler for 5G Networks: Survey and Open Challenges". In: *Advances in Ubiquitous Networking 2*. Springer, 2017, pp. 347–356.

[55] "IEEE Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications". In: *IEEE Std 802.11-2007 (Revision of IEEE Std 802.11-2007)* (2012).

[56] G. Iosifidis et al. "An Iterative Double Auction for Mobile Data Offloading". In: *Proc. of 11th International Symposium on Modeling Optimization in Mobile, Ad Hoc Wireless Networks (WiOpt '13)*. May 2013.

[57] G. Iosifidis et al. "Enabling Crowd-Sourced Mobile Internet Access". In: *Proc. IEEE INFOCOM 2014*. 2014, pp. 451–459.

[58] Xin Kang and Sumei Sun. "Incentive Mechanism Design for Mobile Data Offloading in Heterogeneous Networks". In: *Proc. IEEE International Conference on Communications (ICC) 2015*, pp. 7731–7736.

[59] Chih Heng Ke et al. "A Smart Exponential-Threshold-Linear Backoff Mechanism for IEEE 802.11 WLANs". In: *International Journal of Communication Systems* 24.8 (Aug. 2011), pp. 1033–1048.

[60] S. Kekki et al. *ETSI White Paper No. 28: MEC in 5G networks*. 2018.

[61] Eric Keller et al. "Live migration of an entire network (and its hosts)". In: *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. ACM. 2012, pp. 109–114.

[62] F. P. Kelly, A. K. Maulloo, and D. KH Tan. "Rate control for communication networks: shadow prices, proportional fairness and stability". In: *Journal of the Operational Research society* 49.3 (1998), pp. 237–252.

[63] S. Kurkowski, T. Camp, and M. Colagrosso. "MANET Simulation Studies: The Incredibles". In: *SIGMOBILE Mob. Comput. Commun. Rev.* 9.4 (Oct. 2005), pp. 50–61.

[64] K. Lee et al. "Mobile Data Offloading: How Much Can WiFi Deliver?" In: *Proc. of the 6th ACM International Conference Co-NEXT*. Dec. 2010.

[65] Kyunghan Lee et al. "Mobile Data Offloading: How Much Can WiFi Deliver?" In: *IEEE/ACM Trans. on Networking* 21.2 (Apr. 2013), pp. 536–550.

[66] Chi-Yu Li et al. "Mobile Edge Computing Platform Deployment in 4G LTE Networks: A Middlebox Approach". In: *{USENIX} Workshop on Hot Topics in Edge Computing (HotEdge 18)*. 2018.

[67] Y. Li, D. Papagiannaki, and A. Sheth. "Uplink Traffic Control in Home 802.11 Wireless Networks". In: *Proc. of the 2nd ACM SIGCOMM Workshop on Home Networks*. Aug. 2011.

[68] D. López-Pérez et al. "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments". In: *IEEE Communications Surveys & Tutorials* 17.4 (2015), pp. 2078–2101.

[69] Nguyen Cong Luong et al. "Applications of economic and pricing models for resource management in 5g wireless networks: A survey". In: *IEEE Communications Surveys & Tutorials* (2018).

[70] J. MacKie-Mason, L. Murphy, and J. Murphy. "Responsive pricing in the Internet". In: *Internet Economics* (1997), pp. 279–303.

[71] N. Makris et al. "Cloud-Based Convergence of Heterogeneous RANs in 5G Disaggregated Architectures". In: *IEEE International Conference on Communications (ICC)*. May 2018, pp. 1–6.

[72] N. Makris et al. "Enabling Open Access to LTE network Components; the NITOS testbed paradigm". In: *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*. Apr. 2015, pp. 1–6. DOI: 10.1109/NETSOFT.2015.7116191.

[73] N. Makris et al. "Experimental Evaluation of Functional Splits for 5G Cloud-RANs". In: *IEEE International Conference on Communications (ICC)*. 2017. DOI: 10.1109/ICC.2017.7996493.

[74] Nikos Makris et al. "Employing MEC in the Cloud-RAN: An Experimental Analysis". In: *Proceedings of the 2018 on Technologies for the Wireless Edge Workshop*. New Delhi, India: ACM, 2018, pp. 15–19. ISBN: 978-1-4503-5931-3. DOI: 10.1145/3266276.3266281.

[75] V Miliotis, L Alonso, and C Verikoukis. "Weighted Proportional Fairness and Pricing Based Resource Allocation for Uplink Offloading Using IP Flow Mobility". In: *Ad Hoc Networks* (Oct. 2016).

[76] China Mobile. "C-RAN: the road towards green RAN". In: *White Paper* 2.5 (Oct. 2011).

[77] E. Monsef et al. "Convergence properties of general network selection games". In: *2015 IEEE Conference on Computer Communications (INFOCOM)*. DOI: 10.1109/INFOCOM.2015.7218522.

[78] *Multi-access Edge Computing*. [Online],https://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing.

[79] P. Naghavi et al. "Learning RAT Selection Game in 5G Heterogeneous Networks". In: *IEEE Wireless Communications Letters* 5.1 (2016). ISSN: 2162-2337. DOI: 10.1109/LWC.2015.2495123.

[80] D. D. Nguyen, H. X. Nguyen, and L. B. White. "Performance of adaptive RAT selection algorithms in 5G heterogeneous wireless networks". In: *2016 26th ITNAC*. 2016. DOI: 10.1109/ATNAC.2016.7878785.

[81] A. J. Nicholson and B. D. Noble. "Breadcrumbs: Forecasting Mobile Connectivity". In: *Proc. of the 14th ACM International Conference on Mobile Computing and Networking*. Sept. 2008.

[82] N. Nikaein et al. "OpenAirInterface: A flexible platform for 5G research". In: *ACM SIGCOMM Computer Communication Review* 44.5 (2014), pp. 33–38.

[83] *NITOS - Network Implementation Testbed using Open Source platforms.* [Online],https://nitlab.inf.uth.gr/NITlab/.

[84] D. Niyato and E. Hossain. "Dynamics of Network Selection in Heterogeneous Wireless Networks: An Evolutionary Game Approach". In: *IEEE Transactions on Vehicular Technology* 58.4 (2009). ISSN: 0018-9545. DOI: 10.1109/TVT.2008.2004588.

[85] A. Odlyzko. "Paris Metro Pricing for the Internet". In: *Proc. of the 1st ACM conference on Electronic commerce*. 1999, pp. 140–147.

[86] Maria Oikonomakou et al. "Evaluating Cost Allocation Imposed by Cooperative Switching Off in Multioperator Shared HetNets". In: *IEEE Transactions on Vehicular Technology* 66.12 (2017).

[87] V. Passas et al. "Dynamic RAT Selection and Pricing for Efficient Traffic Allocation in 5G HetNets". In: *Proc. 2019 IEEE ICC*. IEEE. 2019, pp. 1–6.

[88] V. Passas et al. "MATCH: Multiple Access for Multiple Traffic Classes in 5G HetNets". In: *Proc. 2018 IEEE ICC*. IEEE. 2018, pp. 1–6.

[89] V. Passas et al. "Paris Metro Pricing for 5G HetNets". In: *Proc. of IEEE Global Communications Conference (GLOBECOM)*. Dec. 2016.

[90] U. Paul et al. "Understanding Traffic Dynamics in Cellular Data Networks". In: *Proc. of IEEE INFOCOM 2011*. Apr. 2011.

[91] Mugen Peng et al. "Cost-efficient resource allocation in cloud radio access networks with heterogeneous fronthaul expenditures". In: *IEEE Transactions on Wireless Communications* 16.7 (2017), pp. 4626–4638.

[92] Juan S Perez, Sudharman K Jayaweera, and Steven Lane. "Machine learning aided cognitive RAT selection for 5G heterogeneous networks". In: *2017 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*. IEEE. 2017, pp. 1–5.

[93] R. van der Pol et al. "Multipathing with MPTCP and OpenFlow". In: *2012 SC Companion: High Performance Computing, Networking Storage and Analysis*. Nov. 2012, pp. 1617–1624. DOI: 10.1109/SC.Companion.2012.339.

[94] Protobuf-C. *C bindings for Google's Protocol Buffers–Google Project Hosting*.

[95] A. Reznik et al. *ETSI White Paper No. 23: Cloud RAN and MEC: A Perfect Pairing*. 2018.

[96] *RFC 4960: Stream Control Transmission Protocol, 2007*.

[97] *RFC 6824: TCP Extensions for Multipath Operation with Multiple Addresses, 2013*.

[98] CB Sankaran. "Data Offloading Techniques in 3GPP Rel-10 Networks: A tutorial". In: *IEEE Communications Magazine* 50.6 (June 2012), pp. 46–53.

[99] CB Sankaran. "Network access security in next-generation 3GPP systems: A tutorial". In: *IEEE Communications Magazine* 47.2 (2009), pp. 84–91.

[100] O. Sefraoui, M. Aissaoui, and M. Eleuldj. "Article: OpenStack: Toward an Open-source Solution for Cloud Computing". In: *International Journal of Computer Applications* 55.3 (Oct. 2012), pp. 38–42.

[101] S. Sen et al. "A survey of smart data pricing: Past proposals, current plans, and future trends". In: *ACM Computing Surveys (CSUR)* 46.2 (2013), p. 15.

[102] K. Senel and M. Akar. "Fair Resource Allocation in Self-Organizing Heterogeneous Networks with Imperfect Connections". In: *IEEE Transactions on Vehicular Technology* PP.99 (2017), pp. 1–1.

[103] Stephen Soltesz et al. "Container-based Operating System Virtualization: a scalable, high-performance alternative to Hypervisors". In: *ACM SIGOPS Operating Systems Review*. Vol. 41. 3. ACM. 2007, pp. 275–287.

[104] Hyukmin Son et al. "Soft Load Balancing Over Heterogeneous Wireless Networks". In: *IEEE Transactions on Vehicular Technology* 57.4 (July 2008), pp. 2632–2638.

[105] *Speedtest report for Bandwidth Requirements*. URL: https://support.speedtest.net/hc/en-us/articles/203845210-What-speeds-do-I-need-for-Skype-Netflix-video-games-etc-.

[106] T. Stockhammer et al. "MPEG systems technologies - part 6: Dynamic adaptive streaming over HTTP (DASH)". In: ISO/IEC, MPEG Draft International Standard. 2011.

[107] I. Stoica et al. "Chord: A scalable peer-to-peer lookup service for internet applications". In: *ACM SIGCOMM Computer Communication Review* 31.4 (2001), pp. 149–160.

[108] T. Taleb et al. "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration". In: *IEEE Communications Surveys & Tutorials* 19.3 (2017), pp. 1657–1681.

[109] S. Talwar et al. "Enabling Technologies and Architectures for 5G Wireless". In: *Proc. IEEE MTT-S International Microwave Symposium (IMS)*. 2014, pp. 1–4.

[110] *Video DASH dataset*. [Online],http://www-itec.uni-klu.ac.at/ftp/datasets/mmsys12/BigBuckBunny/bunny_1s/.

[111] C. Wang et al. "Cellular architecture and key technologies for 5G wireless communication networks". In: *IEEE Communications Magazine* 52.2 (2014), pp. 122–130.

[112] L. Wang and G. Kuo. "Mathematical Modeling for Network Selection in Heterogeneous Wireless Networks – A tutorial". In: *IEEE Communications Surveys & Tutorials* 15.1 (2013), pp. 271–292.

[113] John Glen Wardrop. "Some Theoretical Aspects of Road Traffic Research". In: *Proc. of the Institution of Civil Engineers* 1 (1952), pp. 325–378.

[114] D. Xenakis et al. "Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms". In: *IEEE Communications Surveys Tutorials* 16.1 (Jan. 2014), pp. 64–91. ISSN: 1553-877X. DOI: 10.1109/SURV.2013.060313.00152.

[115] Wonyong Yoon and Beakcheol Jang. "Enhanced Non-Seamless Offload for LTE and WLAN Networks". In: *IEEE Communications Letters* 17.10 (Oct. 2013), pp. 1960–1963.

[116] Ke Zhang et al. "Energy-efficient offloading for Mobile Edge Computing in 5G Heterogeneous Networks". In: *IEEE access* 4 (2016), pp. 5896–5907.

[117] S. Zhou et al. "The MEC-Based Architecture Design for Low-Latency and Fast Hand-Off Vehicular Networking". In: *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*. 2018, pp. 1–7.

[118] K. Zhu, E. Hossain, and D. Niyato. "Pricing, spectrum sharing, and service selection in two-tier small cell networks: A hierarchical dynamic game approach". In: *IEEE Transactions on Mobile Computing* 13.8 (2014), pp. 1843–1856.