



Predictive and Energy-Aware Management of Heterogeneous Disaggregated 6G RAN

by

Dimitris Kefalas

Supervisors:

Serge Fdida, Professor, Sorbonne University
Thanasis Korakis, Professor, University of Thessaly

Committee:

Thanasis Korakis, Professor, University of Thessaly (Supervisor)
Serge Fdida, Professor, Sorbonne University (Supervisor)
Antonis Argyriou, Professor, University of Thessaly
Fotios Plessas, Professor, University of Thessaly
Paris Flegkas, Assistant Professor, University of Thessaly
Jim Kurose, Professor, University of Massachusetts
Walid Dabbous, Researcher, Inria

June 2026

UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Predictive and Energy-Aware Management of Heterogeneous Disaggregated 6G RAN

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

Dimitris Kefalas

June 2026

UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Gestion prédictive et optimisation énergétique des réseaux d'accès radio 6G désagrégés et hétérogènes

Thèse présentée pour l'obtention du diplôme de Doctorat

Dimitris Kefalas

June 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Έξυπνη και ενεργειακά αποδοτική διαχείριση ετερογενών δικτύων πρόσβασης 6G

Διατριβή η οποία υποβλήθηκε για τη μερική εκπλήρωση
των υποχρεώσεων απόκτησης του Διδακτορικού Διπλώματος

Δημήτρης Κεφαλάς

June 2026





Predictive and Energy-Aware Management of Heterogeneous Disaggregated 6G RAN

PhD Dissertation

Dimitris Kefalas

Advisory committee

Thanasis Korakis, Professor, University of Thessaly (Supervisor)

Serge Fdida, Professor, Sorbonne University (Supervisor)

Examination committee

Antonis Argyriou, Professor, University of Thessaly

Fotios Plessas, Professor, University of Thessaly

Paris Flegkas, Assistant Professor, University of Thessaly

Jim Kurose, Professor, University of Massachusetts

Walid Dabbous, Researcher, Inria

June 2026

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this PH.D. dissertation, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

The declarant

Dimitris Kefalas

Acknowledgments

This PhD has not been just an academic endeavour, but rather a journey that shaped me both as a scientist and, above all, as a person.

My deepest gratitude goes to my supervisor, Prof. Thanasis Korakis, who first believed in me and gave me the opportunity to embark on this journey. I am sincerely thankful for the countless opportunities he has offered me ever since, each of which has helped me grow.

I am equally grateful to my co-supervisor, Prof. Serge Fdida, for the many inspiring scientific discussions we shared. He taught me how to look at a problem from the right angle: not only to identify the questions worth asking, but, more importantly, to find the way to answer them.

I would also like to express my sincere gratitude to the members of my examination committee, Prof. Antonis Argyriou, Prof. Fotios Plessas, Prof. Paris Flegkas, Prof. Jim Kurose, and Dr. Walid Dabbous, for honouring me with their participation, for the time they devoted to reviewing this dissertation, and for their insightful comments and suggestions.

Special thanks go to Dr. Nikos Makris, who stood by me from the very first day of this PhD until the very last. Through his guidance and his remarkable work ethic, he taught me what it truly means to be a good engineer and, through his example, a better professional. Nikos, thank you for being a true mentor to me.

I could not have asked for a better companion on this journey than Sokratis Christakis. From the University of Thessaly to Sorbonne University, we faced the same challenges, shared the same frustrations, and celebrated the same small victories along the way. Of all the things this PhD has given me, the friendship we built is among the ones I value the most.

I am also grateful to the postdoctoral researchers of NITlab, and first of all to Virgilios Passas, as well as to Ilias Syrigos, Kostas Chounos, Kostas Choumas and Apostolis Apostolaras, for their advice and support whenever I needed it. I would also like to thank my colleagues Theodoros Tsourdinis, Dimitris Kanavaris, Stergios Lantzos, Efraim Pavlidis, Mari-

alena Lefkopoulou, Katerina Kyriakou, Giorgos Labis, Alexandros Stoltidis, Ippokratis Koukoulis and many more, for the excellent collaboration and the good moments we shared, both inside and outside the lab. Finally, many thanks to Stavroula Maglavera and the NITlab administrative team, Katerina Arvaniti, Sissy Magouritsa, Olina Stergiopoulou and Stergios Avlogiaris, for always being there to help.

My time in Paris would not have been the same without my colleagues at LIP6, Hassane Rahich, Frederic Vaissade, Taha Mohsen Zaidi, Bilel Zaghoudi and Caterina Leonelli, who welcomed me warmly and made every stay at Sorbonne University a truly enjoyable experience. I am especially thankful to Emilie Mespoulhes, who patiently guided me through every administrative step of the joint PhD and was always willing to help, making life in a new city considerably easier.

Above all, I would like to thank my parents, Giorgos and Pepi, for their incredible support throughout all these years and for their endless love. Nothing I have achieved would have been possible without them, and this PhD belongs to them as much as it does to me. I am also deeply thankful to my sister, Eleftheria, for always being by my side, and to my niece, Katerina, and my nephew who is soon to join us, for filling my life with joy and reminding me what truly matters.

I am also truly grateful to my friends, the people who have walked beside me over these last years and supported me like no one else: Konstantinos Tzalavretas, Spyros Baliotas, Nikos Patsouras, Panagiotis Ganotis, Panagiotis Skoumis, Thodoris Papas, Vasilis Papoutsoglou and Giannis Xaidaras.

A heartfelt thank you also goes to my friends in Paris, and especially to Stavros Baltogiannis, a fellow PhD candidate, for the company and the countless moments we shared throughout my time in France, as well as to Vinia Anastasopoulou, Vaso Maxouri and Fivos De Brito, who made Paris feel like home. I am also grateful to Dr. Andreas Athanasiou for the many fascinating and inspiring conversations we have shared over the years.

Finally, I would like to thank those who played a meaningful part in this journey and supported me immensely along the way, even if life has since taken us in different directions. Your support during those years meant more than words can express, and it will always remain a part of this work.

PhD Dissertation

Predictive and Energy-Aware Management of Heterogeneous Disaggregated 6G RAN

Dimitris Kefalas

Abstract

Fifth-generation (5G) mobile networks introduce cloud-native and disaggregated architectures to support the three distinct service categories of enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC). Looking beyond 5G, 6G evolves these service categories into Immersive Communications (IMC), Hyper-Reliable Low-Latency Communications (HRLLC), and Massive Connectivity (MC), respectively, while introducing three new service categories: Ubiquitous Connectivity (UC) for seamless coverage across heterogeneous and non-terrestrial environments, Integrated AI and Communication (IAC) embedding AI as a native component of the network control plane, and Integrated Sensing and Communication (ISAC) unifying sensing and communication within shared spectrum and hardware. While, 6G defines four design directions: sustainability, connecting the unconnected, security and resilience, and ubiquitous intelligence.

This thesis addresses core challenges in the dynamic control and management of disaggregated 5G and beyond-5G networks, targeting the UC, IMC, HRLLC, and IAC service categories while advancing the sustainability design direction of 6G. Specifically, this thesis addresses the following five research questions: (RQ1) How can non-3GPP access technologies be integrated directly into the disaggregated RAN stack to enable unified capacity extension and service differentiation at the access layer, advancing the UC service category of 6G? (RQ2) Given a heterogeneous multi-RAT RAN resulting from the integration of non-3GPP access technologies, how can AI-driven control mechanisms within the O-RAN architecture enable proactive and intelligent traffic steering across the available RATs, advancing both UC and IAC? (RQ3) As intelligent traffic steering increases the aggregated traffic processed by the RAN, how can cloud-native User Plane functions be autonomously scaled within an O-RAN deployment to sustain service performance under dynamic traffic conditions, target-

ing IMC and HRLLC? (RQ4) Once dynamic RAN User Plane scaling is established, how can different scaling strategies be designed to jointly optimize performance, latency, and energy efficiency in disaggregated O-RAN deployments, advancing HRLLC and the sustainability design direction of 6G? (RQ5) Given that energy-aware scaling relies on knowing the power consumption of RAN components, to what extent can network KPIs predict this consumption for the DU and RU, and how can such predictions support autonomous anomaly detection in online O-RAN deployments, advancing the sustainability design direction and IAC?

Addressing RQ1, a novel architecture is proposed that integrates a non-3GPP WiFi Distributed Unit into a disaggregated 5G C-RAN stack, enabling RAN-level convergence of heterogeneous RATs and advancing the UC service category. A user-based slicing scheme is introduced using SDN and application-aware prioritization within the WiFi MAC layer, enabling per-user QoS differentiation. Experimental validation on a real-world testbed demonstrates improved system capacity and service differentiation without requiring additional licensed spectrum.

Addressing RQ2, the WiFi DU integration is elevated into a fully O-RAN-compliant framework, advancing both UC and IAC. An xApp-based traffic steering mechanism is developed that utilizes LSTM-based demand forecasting within the Near-RT RIC to proactively distribute traffic across heterogeneous, multi-RAT DUs. Evaluation on a real-world testbed demonstrates that the mechanism dynamically adapts to traffic fluctuations, ensuring fair resource distribution and maintaining throughput without degradation under high contention.

Addressing RQ3, a proactive horizontal scaling mechanism for the CU-UP component is proposed, targeting the IMC and HRLLC service categories. An xApp on the Near-RT RIC integrates LSTM-based traffic forecasting to anticipate demand and initiate scaling before congestion occurs. Implementation and evaluation on O-RAN-compliant testbeds using real traffic traces demonstrate enhanced service continuity, reduced packet loss, and more efficient resource utilization compared to static provisioning.

Addressing RQ4, two complementary AI-driven xApps are designed for proactive CU-UP scaling, jointly advancing the sustainability design direction and HRLLC. The first implements vertical scaling by dynamically adjusting CPU resources to minimize power consumption, achieving up to 23% energy savings. The second performs latency-aware horizontal scaling by optimally placing CU-UP instances across the infrastructure, achieving up to 16% energy savings with significantly reduced end-to-end latency.

Addressing RQ5, an AI-driven framework is proposed to advance the sustainability design direction and the IAC by predicting the power consumption of O-DU and O-RU components using only 3GPP-compliant network KPIs. Hybrid XGBoost and CatBoost models achieve a MAPE of 1.02% for the O-RU and 1.22% for the O-DU. Residual-based anomaly detection achieves a 100% detection rate for injected measurement anomalies, followed by automatic calibration.

All proposed mechanisms are designed and evaluated on real-world 5G testbeds under real-time operating conditions, confirming their practical applicability. These five contributions collectively form a path from the physical extension of the RAN through heterogeneous access integration to intelligent and autonomous management of its resources, and ultimately to energy-aware observability of the underlying infrastructure, providing a solid and experimentally validated step toward the autonomous and sustainable RAN architectures envisioned by 6G networks.

Keywords

5G RAN; 5G Multi-RAT; O-RAN architecture; CU-UP scaling; energy-aware orchestration; KPI-driven power prediction; anomaly detection; 6G; Ubiquitous Connectivity; Integrated AI and Communication; sustainability; traffic steering; network slicing

Διδακτορική Διατριβή

Έξυπνη και ενεργειακά αποδοτική διαχείριση ετερογενών δικτύων πρόσβασης 6G

Δημήτρης Κεφαλάς

Περίληψη

Τα δίκτυα κινητής τηλεφωνίας πέμπτης γενιάς (5G) εισάγουν αρχιτεκτονικές τύπου cloud-native και αποσυσχετισμένες (disaggregated), προκειμένου να υποστηρίξουν την Ενισχυμένη Κινητή Ευρυζωνικότητα (ΕΚΕ – eMBB, enhanced Mobile Broadband), την Εξαιρετικά Αξιόπιστη Επικοινωνία Χαμηλής Καθυστέρησης (ΕΑΧΚ – URLLC, Ultra-Reliable Low-Latency Communication) και τη Μαζική Επικοινωνία Μηχανών (ΜΕΜ – mMTC, massive Machine-Type Communication). Πέρα από το 5G, το 6G εξελίσσει αυτές τις κατηγορίες σε Εμβυθιστικές Επικοινωνίες (ΕΕ – IMC, Immersive Communications), Υπερ-Αξιόπιστες Επικοινωνίες Χαμηλής Καθυστέρησης (ΥΑΧΚ – HRLLC, Hyper-Reliable Low-Latency Communications) και Μαζική Συνδεσιμότητα (ΜΣ – MC, Massive Connectivity), αντίστοιχα, ενώ εισάγει τρεις νέες κατηγορίες υπηρεσιών: Πανταχού Παρούσα Συνδεσιμότητα (ΠΠΣ – UC, Ubiquitous Connectivity) για απρόσκοπτη κάλυψη σε ετερογενή και μη επίγεια περιβάλλοντα, Ενσωματωμένη Τεχνητή Νοημοσύνη και Επικοινωνία (ΕΤΝΕ – IAC, Integrated AI and Communication), όπου η Τεχνητή Νοημοσύνη (ΤΝ – AI, Artificial Intelligence) ενσωματώνεται εγγενώς στο επίπεδο ελέγχου του δικτύου, και Ενοποιημένη Αίσθηση και Επικοινωνία (ΕΑΕ – ISAC, Integrated Sensing and Communication), που ενοποιεί την ανίχνευση και την επικοινωνία σε κοινό φάσμα και υλικό. Παράλληλα, το 6G ορίζει τέσσερις βασικές κατευθύνσεις σχεδιασμού: βιωσιμότητα, σύνδεση των μη συνδεδεμένων, ασφάλεια και ανθεκτικότητα, και πανταχού παρούσα νοημοσύνη.

Η παρούσα διατριβή εξετάζει βασικές προκλήσεις στον δυναμικό έλεγχο και τη διαχείριση αποσυσχετισμένων δικτύων 5G και πέραν του 5G, στοχεύοντας στις κατηγορίες υπηρεσιών ΠΠΣ (UC), ΕΕ (IMC), ΥΑΧΚ (HRLLC) και ΕΤΝΕ (IAC), ενώ ταυτόχρονα προάγει τη σχεδιαστική κατεύθυνση της βιωσιμότητας του 6G. Συγκεκριμένα, η διατριβή απαντά στα ακόλουθα πέντε ερευνητικά ερωτήματα: (RQ1) Πώς μπορούν οι τεχνολογίες πρόσβασης εκτός 3GPP (non-3GPP Τεχνολογίες Πρόσβασης Ραδιοεπικοινωνιών – ΤΠΡ, RAT) να

ενσωματωθούν απευθείας στη στοίβα αποσυσχετισμένου Δικτύου Πρόσβασης Ραδιοεπικοινωνιών (ΔΠΡ – RAN, Radio Access Network), ώστε να επιτρέπεται ενοποιημένη επέκταση χωρητικότητας και διαφοροποίηση υπηρεσιών στο επίπεδο πρόσβασης; (RQ2) Πώς μπορούν μηχανισμοί ελέγχου βασισμένοι σε Τεχνητή Νοημοσύνη (ΤΝ – AI) εντός της αρχιτεκτονικής Ανοιχτού Δικτύου Πρόσβασης Ραδιοεπικοινωνιών (ΑΔΠΡ – O-RAN, Open RAN) να επιτρέψουν προληπτική και ευφυή δρομολόγηση κίνησης σε ένα ετερογενές σύστημα πολλαπλών ΤΠΡ (multi-RAT); (RQ3) Πώς μπορούν οι cloud-native λειτουργίες του Επιπέδου Χρήστη (ΕΧ – UP, User Plane) του ΔΠΡ να κλιμακώνονται αυτόνομα σε ένα περιβάλλον ΑΔΠΡ (O-RAN), ώστε να διατηρείται η απόδοση υπηρεσιών υπό δυναμικές και απρόβλεπτες συνθήκες κίνησης; (RQ4) Πώς μπορούν να σχεδιαστούν διαφορετικές στρατηγικές κλιμάκωσης του ΕΧ (UP) ώστε να βελτιστοποιούνται ταυτόχρονα η απόδοση, η καθυστέρηση και η ενεργειακή αποδοτικότητα σε αποσυσχετισμένα περιβάλλοντα ΑΔΠΡ (O-RAN); (RQ5) Σε ποιο βαθμό μπορούν οι Δείκτες Απόδοσης Δικτύου (ΔΑΔ – KPIs, Key Performance Indicators) να προβλέψουν την κατανάλωση ισχύος των στοιχείων του αποσυσχετισμένου ΔΠΡ και πώς μπορεί αυτή η σχέση να υποστηρίξει αυτόνομη ανίχνευση ανωμαλιών σε online αναπτύξεις ΑΔΠΡ (O-RAN);

Αντιμετωπίζοντας το RQ1, προτείνεται μια νέα αρχιτεκτονική που ενσωματώνει μια Κατανεμημένη Μονάδα WiFi (ΚΜ – WiFi DU, Distributed Unit) εκτός 3GPP σε μια αποσυσχετισμένη στοίβα 5G ΔΠΡ Νέφους (ΔΠΡΝ – C-RAN, Cloud RAN), επιτρέποντας σύγκλιση ετερογενών ΤΠΡ σε επίπεδο ΔΠΡ και προωθώντας την κατηγορία ΠΠΣ (UC). Εισάγεται ένα σχήμα τεμαχισμού ανά χρήστη (user-based slicing), χρησιμοποιώντας Δικτύωση Οριζόμενη από Λογισμικό (ΔΟΛ – SDN, Software-Defined Networking) και προτεραιοποίηση με επίγνωση εφαρμογής στο επίπεδο Ελέγχου Πρόσβασης Μέσου (ΕΠΜ – MAC, Medium Access Control) του WiFi, επιτρέποντας διαφοροποίηση Ποιότητας Υπηρεσίας (ΠΥ – QoS, Quality of Service) ανά χρήστη. Πειραματική αξιολόγηση σε πραγματικό testbed δείχνει βελτιωμένη χωρητικότητα συστήματος και διαφοροποίηση υπηρεσιών χωρίς ανάγκη επιπλέον αδειοδοτημένου φάσματος.

Αντιμετωπίζοντας το RQ2, η ενσωμάτωση του WiFi DU επεκτείνεται σε πλήρως συμβατό πλαίσιο ΑΔΠΡ (O-RAN), ενισχύοντας τις κατηγορίες ΠΠΣ (UC) και ΕΤΝΕ (IAC). Αναπτύσσεται ένας μηχανισμός δρομολόγησης κίνησης βασισμένος σε xApp, αξιοποιώντας πρόβλεψη ζήτησης με Μακρά Βραχυπρόθεσμη Μνήμη (ΜΒΜ – LSTM, Long Short-Term Memory) εντός του Ελεγκτή Ευφυΐας ΔΠΡ Σχεδόν Πραγματικού Χρόνου (ΕΕΔΠΡ-ΣΠΧ –

Near-RT RIC, Near Real-Time RAN Intelligent Controller), ώστε να κατανέμεται προληπτικά η κίνηση μεταξύ ετερογενών ΚΜ (DU). Η αξιολόγηση σε πραγματικό testbed δείχνει ότι ο μηχανισμός προσαρμόζεται δυναμικά στις διακυμάνσεις της κίνησης, εξασφαλίζοντας δίκαιη κατανομή πόρων και διατηρώντας την απόδοση χωρίς υποβάθμιση υπό υψηλό ανταγωνισμό.

Αντιμετωπίζοντας το RQ3, προτείνεται ένας μηχανισμός προληπτικής οριζόντιας κλιμάκωσης για το στοιχείο Κεντρικής Μονάδας – Επιπέδου Χρήστη (ΚΜ-ΕΧ – CU-UP, Central Unit – User Plane), στοχεύοντας τις κατηγορίες ΕΕ (IMC) και ΥΑΧΚ (HRLLC). Ένα xApp στον ΕΕΔΠΡ-ΣΠΧ (Near-RT RIC) ενσωματώνει πρόβλεψη κίνησης με ΜΒΜ (LSTM) ώστε να προβλέπει τη ζήτηση και να ενεργοποιεί την κλιμάκωση πριν την εμφάνιση συμφόρησης. Η υλοποίηση και αξιολόγηση σε testbeds συμβατά με ΑΔΠΡ (O-RAN), χρησιμοποιώντας πραγματικά ίχνη κίνησης, δείχνουν βελτιωμένη συνέχεια υπηρεσίας, μειωμένη απώλεια πακέτων και αποδοτικότερη χρήση πόρων σε σύγκριση με στατικές προσεγγίσεις.

Αντιμετωπίζοντας το RQ4, σχεδιάζονται δύο συμπληρωματικά xApps βασισμένα σε ΤΝ για προληπτική κλιμάκωση του ΚΜ-ΕΧ (CU-UP), προωθώντας τόσο τη βιωσιμότητα όσο και την ΥΑΧΚ (HRLLC). Το πρώτο υλοποιεί κατακόρυφη κλιμάκωση μέσω δυναμικής προσαρμογής πόρων CPU, επιτυγχάνοντας έως 23% εξοικονόμηση ενέργειας. Το δεύτερο υλοποιεί οριζόντια κλιμάκωση με επίγνωση καθυστέρησης, τοποθετώντας βέλτιστα τα στιγμιότυπα CU-UP στην υποδομή, επιτυγχάνοντας έως 16% εξοικονόμηση ενέργειας και σημαντική μείωση της end-to-end καθυστέρησης.

Αντιμετωπίζοντας το RQ5, προτείνεται ένα πλαίσιο βασισμένο σε ΤΝ που προωθεί τη βιωσιμότητα και την ΕΤΝΕ (IAC), προβλέποντας την κατανάλωση ισχύος των στοιχείων Κατανεμημένης Μονάδας (ΚΜ – O-DU, O-RAN Distributed Unit) και Μονάδας Ραδιοεπικοινωνίας (ΜΡ – O-RU, O-RAN Radio Unit) χρησιμοποιώντας αποκλειστικά ΔΑΔ (KPIs) συμβατούς με 3GPP. Υβριδικά μοντέλα XGBoost και CatBoost επιτυγχάνουν Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (ΜΑΠΣ – MAPE, Mean Absolute Percentage Error) 1,02% για το O-RU και 1,22% για το O-DU. Η ανίχνευση ανωμαλιών βάσει υπολοίπων επιτυγχάνει ποσοστό ανίχνευσης 100% για τεχνητά εισαγόμενες ανωμαλίες με επακόλουθη αυτόματη βαθμονόμηση.

Όλοι οι προτεινόμενοι μηχανισμοί σχεδιάστηκαν και αξιολογήθηκαν σε πραγματικά testbeds 5G υπό συνθήκες λειτουργίας πραγματικού χρόνου, επιβεβαιώνοντας την πρακτική τους εφαρμοσιμότητα. Οι πέντε αυτές συνεισφορές διαμορφώνουν μια συνεκτική πορεία από τη

φυσική επέκταση του ΔΠΡ μέσω ετερογενούς πρόσβασης, στην ευφυή και αυτόνομη διαχείριση πόρων, και τελικά στην ενεργειακά αποδοτική παρατηρησιμότητα της υποκείμενης υποδομής, αποτελώντας ένα ουσιαστικό και πειραματικά τεκμηριωμένο βήμα προς τα αυτόνομα και βιώσιμα ΔΠΡ που οραματίζεται το 6G.

Keywords

5G RAN; O-RAN architecture; CU-UP scaling; hybrid access networks; energy-aware orchestration

Thèse de Doctorat

Gestion prédictive et optimisation énergétique des réseaux d'accès radio 6G désagrégés et hétérogènes

Dimitris Kefalas

Résumé

Les réseaux mobiles de cinquième génération (5G) introduisent des architectures cloud-natives et désagrégées pour prendre en charge les trois catégories de services distinctes que sont le haut débit mobile amélioré (eMBB), les communications ultra-fiables à faible latence (URLLC) et les communications massives de type machine (mMTC). Au-delà de la 5G, la 6G fait évoluer ces catégories de services respectivement en Communications Immersives (IMC), Communications Hyper-Fiables à Faible Latence (HRLLC) et Connectivité Massive (MC), tout en introduisant trois nouvelles catégories de services : la Connectivité Ubiquitaire (UC) pour une couverture transparente à travers des environnements hétérogènes et non terrestres, l'Intelligence Artificielle et Communication Intégrées (IAC) intégrant l'IA comme composant natif du plan de contrôle du réseau, et la Détection et Communication Intégrées (ISAC) unifiant détection et communication au sein d'un spectre et d'un matériel partagés. Par ailleurs, la 6G définit quatre axes de conception : la durabilité, la connexion des non-connectés, la sécurité et la résilience, et l'intelligence ubiquitaire. Cette thèse aborde les défis fondamentaux liés au contrôle dynamique et à la gestion des réseaux 5G désagrégés et au-delà de la 5G, en ciblant les catégories de services UC, IMC, HRLLC et IAC tout en faisant progresser l'axe de conception « durabilité » de la 6G. Plus précisément, cette thèse traite les cinq questions de recherche suivantes : (RQ1) Comment les technologies d'accès non-3GPP peuvent-elles être intégrées directement dans la pile RAN désagrégée afin de permettre une extension unifiée de la capacité et une différenciation des services au niveau de la couche d'accès, faisant ainsi progresser la catégorie de service UC de la 6G ? (RQ2) Étant donné un RAN hétérogène multi-RAT résultant de l'intégration de technologies d'accès non-3GPP, comment des mécanismes de contrôle pilotés par l'IA au sein de l'architecture O-RAN peuvent-ils permettre un pilotage proactif et intelligent du trafic à travers les RAT disponibles, faisant progresser à la fois l'UC et l'IAC ? (RQ3) Alors que le pi-

lotage intelligent du trafic augmente le trafic agrégé traité par le RAN, comment les fonctions cloud-natives du Plan Utilisateur peuvent-elles être dimensionnées de manière autonome au sein d'un déploiement O-RAN afin de maintenir les performances de service dans des conditions de trafic dynamiques et imprévisibles, ciblant l'IMC et l'HRLLC ? (RQ4) Une fois le dimensionnement dynamique du Plan Utilisateur établi, comment différentes stratégies de dimensionnement peuvent-elles être conçues pour optimiser conjointement les performances, la latence et l'efficacité énergétique dans les déploiements O-RAN désagrégés, faisant progresser l'HRLLC et l'axe de conception « durabilité » de la 6G ? (RQ5) Étant donné que le dimensionnement tenant compte de l'énergie repose sur la connaissance de la consommation électrique des composants RAN, dans quelle mesure les KPI réseau peuvent-ils prédire cette consommation pour le DU et le RU, et comment de telles prédictions peuvent-elles soutenir la détection autonome d'anomalies dans les déploiements O-RAN en ligne, faisant progresser l'axe de conception « durabilité » et l'IAC ? En réponse à RQ1, une architecture novatrice est proposée, intégrant une Unité Distribuée WiFi non-3GPP dans une pile C-RAN 5G désagrégée, permettant la convergence au niveau RAN de RAT hétérogènes et faisant progresser la catégorie de service UC. Un schéma de découpage par utilisateur est introduit à l'aide du SDN et d'une priorisation tenant compte des applications au sein de la couche MAC WiFi, permettant une différenciation de la QoS par utilisateur. La validation expérimentale sur un banc d'essai réel démontre une amélioration de la capacité du système et de la différenciation des services sans nécessiter de spectre sous licence supplémentaire. En réponse à RQ2, l'intégration du DU WiFi est élevée au rang de cadre pleinement conforme à O-RAN, faisant progresser à la fois l'UC et l'IAC. Un mécanisme de pilotage du trafic basé sur une xApp est développé, utilisant la prévision de la demande par LSTM au sein du Near-RT RIC pour distribuer proactivement le trafic à travers les DU multi-RAT hétérogènes. L'évaluation sur un banc d'essai réel démontre que le mécanisme s'adapte dynamiquement aux fluctuations du trafic, assurant une distribution équitable des ressources et maintenant le débit sans dégradation en cas de forte contention. En réponse à RQ3, un mécanisme proactif de dimensionnement horizontal du composant CU-UP est proposé, ciblant les catégories de services IMC et HRLLC. Une xApp sur le Near-RT RIC intègre la prévision du trafic par LSTM pour anticiper la demande et initier le dimensionnement avant l'apparition de la congestion. L'implémentation et l'évaluation sur des bancs d'essai conformes à O-RAN, utilisant des traces de trafic réelles, démontrent une amélioration de la continuité de service,

une réduction des pertes de paquets et une utilisation plus efficace des ressources par rapport au provisionnement statique. En réponse à RQ4, deux xApps complémentaires pilotées par l'IA sont conçues pour le dimensionnement proactif du CU-UP, faisant progresser conjointement l'axe de conception « durabilité » et l'HRLLC. La première met en œuvre un dimensionnement vertical en ajustant dynamiquement les ressources CPU pour minimiser la consommation électrique, atteignant jusqu'à 23 % d'économies d'énergie. La seconde réalise un dimensionnement horizontal tenant compte de la latence en plaçant de manière optimale les instances CU-UP à travers l'infrastructure, atteignant jusqu'à 16 % d'économies d'énergie avec une latence de bout en bout significativement réduite. En réponse à RQ5, un cadre piloté par l'IA est proposé, faisant progresser l'axe de conception « durabilité » et l'IAC en prédisant la consommation électrique des composants O-DU et O-RU à l'aide uniquement de KPI réseau conformes au 3GPP. Des modèles hybrides XGBoost et CatBoost atteignent un MAPE de 1,02% pour l'O-RU et de 1,22% pour l'O-DU. La détection d'anomalies basée sur les résidus atteint un taux de détection de 100% pour les anomalies de mesure injectées, suivie d'un calibrage automatique. Tous les mécanismes proposés sont conçus et évalués sur des bancs d'essai 5G réels dans des conditions de fonctionnement en temps réel, confirmant leur applicabilité pratique. Ces cinq contributions forment collectivement un cheminement allant de l'extension physique du RAN par l'intégration d'accès hétérogènes, à la gestion intelligente et autonome de ses ressources, et enfin à l'observabilité tenant compte de l'énergie de l'infrastructure sous-jacente, fournissant une avancée solide et expérimentalement validée vers les architectures RAN autonomes et durables que les réseaux 6G envisagent.

Keywords

5G RAN; O-RAN architecture; CU-UP scaling; hybrid access networks; energy-aware orchestration

List of Publications

Journals

[J1] D. Kefalas, N. Makris, S. Fdida and T. Korakis, **Proactive Scaling and Energy Optimization for 5G O-RAN: Experimental Insights**, IEEE Open Journal of the Communications Society, vol. 7, pp. 10626-10644, 2026. [1]

Conferences

[C1] D. Kefalas, N. Makris, T. Korakis and S. Fdida, **Design, Integration and Slicing of non-3GPP DUs in 5G Networks: a Hybrid 5G-DU Approach**, IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Vancouver, BC, Canada, 2024. [2]

[C2] D. Kefalas, N. Makris, T. Korakis and S. Fdida, **Traffic Steering for O-RAN Multi-RAT 5G Networks Using ML-Based Demand Forecasting**, 2025 IEEE Conference on Network Function Virtualization and Software-Defined Networking (NFV-SDN), Athens, Greece, 2025. [3]

[C3] D. Kefalas, N. Makris, S. Fdida and T. Korakis, **Enhancing 5G V-RAN Reliability via Reactive CU-UP Scaling**, 2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Poznan, Poland, 2025. [4]

[C4] D. Kefalas, N. Makris, S. Fdida and T. Korakis, **On the Feasibility of RAN Scaling for Beyond 5G Networks: Proactive CU-UP Scaling with ML Demand Forecasting**, 2025 IEEE 11th International Conference on Network Softwarization (NetSoft), Budapest, Hungary, 2025. [5]

[C5] D. Kefalas, N. Makris, S. Fdida, T. Korakis, **Prediction, Anomaly Detection and Calibration of O-RAN Power Consumption: An AI-Driven Approach Using Network KPIs**. Fifth International Workshop on Green and Sustainable Networking 2026 (GreenNet 2026), 2026. [6]

In addition to the core research presented in this volume, the author co-authored the following works.

Conferences

[C6] I. Syrigos, D. Kefalas, N. Makris and T. Korakis, **EELAS: Energy Efficient and Latency Aware Scheduling of Cloud-Native ML Workloads**, 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2023. [7]

[C7] D. Kefalas, S. Christakis, S. Fdida, N. Makris; I. Syrigos; V. Passas, and T. Korakis, **slAIces: an LLM Chatbot for Simplifying Experiments with the SLICES-RI**, 2024 IFIP Networking Conference (IFIP Networking), Thessaloniki, Greece, 2024. [8]

[C8] C. Leonelli, D. Kefalas, S. Fdida, P. Bellavista and T. Korakis, **Dynamic Resource Allocation and Energy Optimization in 5G O-RAN: Real-World Insights and Testbed Evaluations**, 2025 IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, QC, Canada, 2025 [9]

[C9] S. Christakis, D. Kefalas, N. Makris, S. Fdida and T. Korakis, **Toward Multi-RAT 6G Networks: Real-Radio Evaluation of Multipath Transport over a 5G ATSSS Testbed**, 2026 IEEE Future Networks World Forum (FNWF), 2026 (accepted, to appear).

Table of contents

Abstract **xiii**

Περίληψη **xvii**

Résumé **xxi**

List of Publications **xxv**

Table of contents **xxvii**

List of figures **xxxiii**

List of tables **xxxvii**

Abbreviations **xxxix**

1 Introduction **1**

1.1 From 5G to 6G: An Architectural Evolution 1

1.2 Open Research Challenges in 6G RAN Control and Management 4

1.3 Thesis Contributions to 6G Network Management 7

1.3.1 Enabling Multi-RAT Access via WiFi DU Integration in 5G RAN . 7

1.3.2 Proactive, ML-Driven Traffic Steering in Multi-RAT O-RAN Systems 7

1.3.3 On the feasibility of Dynamic Scaling of the 5G O-RAN User Plane 8

1.3.4 Energy-Efficient Proactive CU-UP Scaling in 5G O-RAN: Vertical and Latency-Aware Horizontal Strategies 8

1.3.5 KPI-Driven Power Consumption Prediction and Anomaly Detection in O-RAN Networks 9

1.4 Thesis Structure 10

2 Background **13**

2.1 From analog to digital Cellular networks 13

2.2 5G Architecture 14

2.2.1 5G NR Protocol Stack Overview 16

2.2.2 Disaggregated and Cloud-Native Network Architectures 17

2.3 The O-RAN AI/ML Workflow 21

2.4 Non-3GPP Access Technologies 21

2.4.1 IEEE 802.11 - WiFi 22

2.5 Deployment Tools 23

2.5.1 5G software suites 24

2.5.2 WiFi Access Point Software 24

2.5.3 Containers 25

2.5.4 Kubernetes 25

2.6 Testbeds 26

2.6.1 SLICES Research Infrastructure (SLICES RI) 26

2.6.2 5GTNF 26

2.7 Dataset 27

2.8 Conclusion 28

3 Enabling Multi-RAT Access via WiFi DU Integration in 5G RAN **29**

3.1 Introduction 29

3.2 Related Work 31

3.3 System Model and Architecture 33

3.3.1 WiFi DU integration 33

3.3.2 WiFi Slicing Mechanism 34

3.4 Evaluation 36

3.4.1 Experiment Setup 36

3.4.2 Experiments 38

3.5 Conclusions 41

4 Proactive, ML-Driven Traffic Steering in Multi-RAT O-RAN Systems **43**

4.1 Introduction 43

4.2 Related Work 45

4.3 System Architecture and Environment Setup 46
4.3.1 System Architecture 47
4.3.2 xApp Functionality 48
4.3.3 The Necessity of ML-Based Demand Forecasting 50
4.3.4 Real World Implementation 50
4.4 Experimental Environment & Evaluation 52
4.4.1 Experimental Conditions 52
4.4.2 Experiments & Evaluation 53
4.5 Conclusion 55

5 On the feasibility of Dynamic Scaling of the 5G O-RAN User Plane 57
5.1 Introduction 57
5.2 Related Work 59
5.3 Empirical Analysis of CU-UP CPU Utilization and Traffic Load 61
5.4 System Model and Architecture 62
5.4.1 5G Network Configuration and Setup 63
5.4.2 Horizontal CU-UP Scaling Process 66
5.4.3 Machine Learning Model - Demand Prediction 68
5.4.4 xApp Functionality 69
5.4.5 Experimentation environment 69
5.5 Evaluation 71
5.5.1 Dataset 71
5.5.2 Experiments 71
5.6 Discussion & Limitations 74
5.7 Conclusion 75

6 Energy-Efficient Proactive CU-UP Scaling in 5G O-RAN: Vertical and Latency-Aware Horizontal Strategies 77
6.1 Introduction 77
6.2 Related Work 79
6.3 CU-UP Power Consumption Modelling 81
6.4 System Architecture 84
6.4.1 Machine Learning Model 85

6.4.2 Vertical Scaling . . . 88
6.4.3 Horizontal Scaling . . . 88
6.5 Experiments . . . 95
6.5.1 5G Network Configuration and Setup . . . 95
6.5.2 Real-World Traffic Generation . . . 97
6.5.3 CU-UP scaling process . . . 98
6.5.4 Evaluation . . . 100
6.6 Discussion . . . 108
6.7 Conclusion . . . 109

7 KPI-Driven Power Consumption Prediction and Anomaly Detection in O-RAN Networks 111
7.1 Introduction . . . 111
7.2 Related Work . . . 113
7.3 System Architecture . . . 114
7.4 Methodology & AI Model Design . . . 117
7.4.1 Dataset . . . 117
7.4.2 Correlation Analysis . . . 117
7.4.3 Machine Learning Models . . . 119
7.5 Experimental Results & Validation . . . 121
7.6 Conclusion . . . 124

8 Conclusion 125
8.1 Summary of Contributions . . . 125
8.2 Perspectives for Future Work . . . 128
8.2.1 Towards Full Multi-RAT Orchestration with Non-Terrestrial Networks 128
8.2.2 Hybrid and Cross-Layer RAN Scaling Strategies . . . 128
8.2.3 Energy Management Across All O-RAN Components . . . 129
8.2.4 Foundation Models and Generative AI for RAN Management . . . 129
8.2.5 Towards a Joint RAN and AI Workload Scheduling for AI-RAN . . . 130
8.3 Lessons Learned . . . 131
8.3.1 Real-World Testbed Experimentation as a Foundation for Applicability 131
8.3.2 O-RAN as a Foundation for Self-Autonomous Networks . . . 132

8.3.3 Native AI Integration as a Design Principle for 6G RAN Management 132
8.4 Outlook 133

References **135**

List of figures

1.1 6G Service Categories . . . 3

2.1 5G NSA vs SA architectures . . . 15
2.2 5G SA CN . . . 16
2.3 Overview of 5G NR Protocol Stack CP & UP . . . 17
2.4 RAN Functional Splits . . . 18
2.5 O-RAN architecture . . . 20
2.6 WLAN architecture . . . 23

3.1 5G Disaggregated Architecture . . . 37
3.2 WiFi DU slicing architecture . . . 38
3.3 RTT of packets across multiple DUs technologies . . . 39
3.4 DL UDP traffic across multiple DUs technologies . . . 39
3.5 UDP transmission to two stations without slicing . . . 39
3.6 UDP transmission to two stations slicing supported . . . 39

4.1 Proactive scaling O-RAN system architecture . . . 46
4.2 UE1 Throughput . . . 53
4.3 UE2 Throughput . . . 53
4.4 UE1 - WiFi DU Utilization . . . 53
4.5 UE2 - WiFi DU Utilization . . . 53
4.6 UE1 - 5G DU Utilization . . . 53
4.7 UE2 - 5G DU Utilization . . . 53

5.1 Correlation Between Incoming Traffic Load and CU-UP CPU Utilization . . . 61
5.2 Downlink Traffic Profile Driving CU-UP Processing Demand . . . 62
5.3 CU-UP CPU Utilization . . . 62

5.4 Proactive scaling O-RAN system architecture 63
5.5 Experimental 5G CN and O-RAN architecture 65
5.6 Naive System: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate. . . 70
5.7 Reactive CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate. 70
5.8 Proactive CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate. 70

6.1 Impact of CPU Resources on CU-UP Power Consumption vs Bandwidth . . 81
6.2 Regression Models for CU-UP power consumption 84
6.3 5G O RAN Architecture Supporting CU UP Vertical or Horizontal Scaling Through Dedicated xApps 85
6.4 ML models prediction architecture 86
6.5 Experimental Architecture 98
6.6 Proactive Vertical CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate. 99
6.7 CU-UP Power Consumption Vertical Scaling vs Static Overprovisioning Across Sessions 99
6.8 CU-UP Power Saving of Vertical Scaling against Static Overprovisioning . 100
6.9 Power Consumption of Vertical Scaling xApp 100
6.10 Proactive CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate. 103
6.11 CU-UP Power Consumption Horizontal Scaling vs Static Overprovisioning Across Sessions 104
6.12 CU-UP Power Saving of Horizontal Scaling against Static Overprovisioning 105
6.13 Power Consumption of Horizontal Scaling xApp 105
6.14 CU-UP Power Saving Comparison between Vertical and Horizontal Scaling System 106
6.15 UE's Mean RTT under Different Scaling Configurations 107

7.1 System Architecture and O-RAN Testbed Deployment 115
7.2 Residual-Based Power Anomaly Detection Architecture 116
7.3 Generated DL DataRate during experiments 120

7.4 RU Power Forecast vs. Measured Power 122
7.5 Detection of Injected +8% Power Anomalies on the RU Node 122
7.6 DU Power Forecast vs. Measured Power 123
7.7 Detection of Injected +8% Power Anomalies on the DU Node 123

List of tables

2.1 Dataset KPIs 28

3.1 Mapping of WiFi AC and MAC Queues 35
3.2 ToS-UP mapping 35

5.1 Node Assignments and Specifications 65

6.1 Comparison of state-of-the-art with our approach. 82
6.2 Node Assignments and Specifications 97
6.3 RTT Latency Between CU-UP and DU Nodes 97
6.4 Maximum Sustainable Throughput per CU-UP CPU Provisioning Profile . 100
6.5 Energy Consumption and Savings Over the Whole Execution 106
6.6 Energy Consumption and Savings During the Non-Overprovisioned Period 106
6.7 Energy Consumption per Traffic Regime: Vertical vs. Horizontal Scaling . . 107

7.1 Comparison of DU and RU KPIs Correlation with Power Consumption (W) 117
7.2 Comparison of Top 5 Performance Models for DU and RU Components . . 121

Abbreviations

3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
5GCN	5G Core Network
5GTN	5G Test Network
6G	6th Generation
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
AP	Access Point
ARIMA	Autoregressive Integrated Moving Average
ATSSS	Access Traffic Steering, Switching, and Splitting
AUSF	Authentication Server Function
AWGN	Additive White Gaussian Noise
B5G	Beyond 5G
BAAS	Bit Rate Aware Auto Scaling
BBU	Baseband Unit
BE	Best Effort
BK	Background
BS	Base Station
CN	Core Network
CNN	Convolutional Neural Network
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPU	Central Processing Unit
CQI	Channel Quality Indicator

C-RAN	Cloud Radio Access Network
CSMA	Carrier Sense Multiple Access
CU	Centralized Unit
CU-CP	CU Control Plane
CUPS	Control and User Plane Separation
CU-UP	CU User Plane
DCT	Digital Communication Technologies
DL	Downlink
DNN	Deep Neural Network
D-RAN	Distributed Radio Access Network
DRB	Data Radio Bearer
DRL	Deep Reinforcement Learning
DSCP	Differentiated Services Code Point
DU	Distributed Unit
E1AP	E1 Application Protocol
E2E	End-to-End
E2SM-KPM	E2 Service Model for Key Performance Measurement
eMBB	Enhanced Mobile Broadband
eNB	Evolved Node B
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
F1AP	F1 Application Protocol
FH	Fronthaul
FLP	Facility Location Problem
gNB	Next Generation Node B
GTP-U	GPRS Tunnelling Protocol User Plane
HARQ	Hybrid Automatic Repeat Request
HPA	Horizontal Pod Autoscaler
HPC	High Performance Computing
HRLLC	Hyper-Reliable Low-Latency Communications
HSS	Home Subscriber Server
HT	High Throughput

IAC	Integrated AI and Communication
ICMP	Internet Control Message Protocol
IMC	Immersive Communications
IMT	International Mobile Telecommunications
IoT	Internet of Things
ISAC	Integrated Sensing and Communication
ITU-R	International Telecommunication Union – Radiocommunication Sector
KPI	Key Performance Indicator
LDPC	Low-Density Parity-Check
LGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
LTE	Long Term Evolution
LWA	LTE-WLAN Aggregation
LWIP	LTE-WLAN Radio Level Integration with IPsec Tunnel
MAC	Medium Access Control
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MC	Massive Connectivity
MCS	Modulation and Coding Scheme
MEC	Multi-access Edge Computing
ML	Machine Learning
MME	Mobility Management Entity
mMTC	Massive Machine-Type Communications
MPTCP	Multipath TCP
MU-MIMO	Multi-User Multiple-Input Multiple-Output
N3IWF	Non-3GPP Interworking Function
NFV	Network Function Virtualization
NG-RAN	Next Generation Radio Access Network
NR	New Radio
NRF	Network Repository Function
NSA	Non-Standalone
NSSF	Network Slice Selection Function

NTN	Non-Terrestrial Network
OAI	OpenAirInterface
O-CU	O-RAN Centralized Unit
O-DU	O-RAN Distributed Unit
OFDM	Orthogonal Frequency-Division Multiplexing
O-RAN	Open Radio Access Network
O-RU	O-RAN Radio Unit
OSA	OpenAirInterface Software Alliance
PA	Power Amplifier
PCF	Policy and Charging Function
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
P-GW	Packet Data Network Gateway
PHY	Physical Layer
PRB	Physical Resource Block
PUCCH	Physical Uplink Control Channel
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RAN	Radio Access Network
RAPL	Running Average Power Limit
rApp	Non-RT RIC Application
RAT	Radio Access Technology
RC	Radio Config
RF	Radio Frequency
RIC	RAN Intelligent Controller
RLC	Radio Link Control
RMSE	Root Mean Square Error
RNTI	Radio Network Temporary Identifier
RRC	Radio Resource Control
RRU	Remote Radio Unit
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality

RSSI	Received Signal Strength Indicator
RTT	Round-Trip Time
RU	Radio Unit
SA	Standalone
SBA	Service-Based Architecture
SDAP	Service Data Adaptation Protocol
SDN	Software-Defined Networking
S-GW	Serving Gateway
SLA	Service Level Agreement
SLICES-RI	Scientific Large-Scale Infrastructure for Computing/Communication Experimental Stu
SMF	Session Management Function
SMO	Service Management and Orchestration
SNR	Signal-to-Noise Ratio
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TNGF	Trusted Non-3GPP Gateway Function
TTI	Transmission Time Interval
UC	Ubiquitous Connectivity
UDM	Unified Data Management
UDP	User Datagram Protocol
UDR	Unified Data Repository
UE	User Equipment
UL	Uplink
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low-Latency Communications
USRP	Universal Software Radio Peripheral
VHT	Very High Throughput
VNF	Virtual Network Function
VO	Voice
VR	Virtual Reality
V-RAN	Virtualized Radio Access Network

WLAN	Wireless Local Area Network
WMM	WiFi Multimedia
xApp	Near-RT RIC Application
XR	Extended Reality

Chapter 1

Introduction

1.1 From 5G to 6G: An Architectural Evolution

Cellular networks have evolved from circuit-switched voice systems to software-driven architectures. This shift culminates in 5th-generation cellular networks (5G), where the disaggregated and cloud-native Radio Access Network (RAN), alongside the Open RAN (O-RAN) framework, introduces unprecedented flexibility in the deployment and real-time control of radio resources. 6G aims to create a fully autonomous and intelligence-based cellular network system that incorporates Artificial Intelligence (AI) as a core architectural principle [10], [11], rather than a simple supporting tool as used in the 5G era.

5G networks support three distinct categories of services: enhanced mobile broadband (eMBB), Ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC). The target of the eMBB category is to offer maximum data transfer rates and high spectral efficiency for applications requiring high bandwidth, while the URLLC was introduced for mission-critical services that require extremely low latency. Regarding the mMTC, this service aims to offer large-scale connectivity of billions of devices. In order to meet the requirements of the aforementioned services within a unified network infrastructure, 5G departed from the monolithic architectures in the Core Network (CN) and RAN that were dominating previous generations of the cellular networks. Instead, it introduced a service-based architecture in the CN and a disaggregated architecture in the RAN, in which the base station is split into a Radio Unit (RU), Distributed Unit (DU), and Centralized Unit (CU), as standardized by the 3GPP [12]. The RAN evolved from a hardware-based system to a software-defined and virtualized infrastructure deployable on commercial off-

the-shelf (COTS) hardware. Building on the foundations laid by RAN disaggregation, a new concept introduced by the O-RAN Alliance further expanded this vision by introducing open interfaces and the RAN Intelligent Controller (RIC), enabling the intelligent control of RAN components through the control-level applications running on top of the 5G RAN [10]. This has led to architectures that create the foundation for flexible and intelligent network management at the RAN level.

Despite these developments, both real-world 5G deployments and the future applications that Beyond 5G (B5G) networks are expected to support (e.g., Massive IoT and Industry 4.0, Immersive and Extended Reality, etc.) reveal significant operational limitations that are driving the transition to 6G [13]. Some of 5G's limitations can be summarized as follows. The observed latency of real-world 5G deployments ranges above 5 ms, making it impossible to support critical applications such as telesurgery, collaborative robots, connected & autonomous vehicles and others that require less than 1 ms [14]. The maximum throughput of ≈ 10 Gbps that 5G offers under ideal conditions is insufficient for various future applications (e.g., holographic video and immersive XR), which require Tbps-level bandwidth [15]. Additionally, the mmWave frequencies used to achieve these data rates pose serious coverage challenges, with limited range in indoor environments and limited ability to penetrate obstacles [16]. Apart from performance, the management of disaggregated cloud-native RAN components remains largely static or reactive, unable to adapt efficiently to the dynamic and heterogeneous traffic demands of online networks, resulting in congestion under peak load or unnecessary resource overprovisioning during periods of low usage [17]. At the same time, operating 5G networks consumes enormous amounts of energy. Numerically speaking, the 5G BS, the gNB, consumes up to 3 times as much energy as the 4G eNB BS [18], since energy efficiency was not a primary design objective in 5G, raising critical sustainability concerns for the future of cellular networks.

Looking beyond 5G, 6G, as envisioned by the ITU-R IMT-2030 framework [19], aims to address the aforementioned limitations. 6G evolves the eMBB service category into immersive communications (IMC), extending beyond the high bandwidth capabilities in order to support the vast data needs of next-generation applications such as holographic displays and haptic feedback. Regarding the mMTC, it has evolved into massive connectivity (MC), offering connectivity to an even more heterogeneous ecosystem of connected entities, such as intelligent sensors, nano-devices, non-terrestrial network nodes, and others.

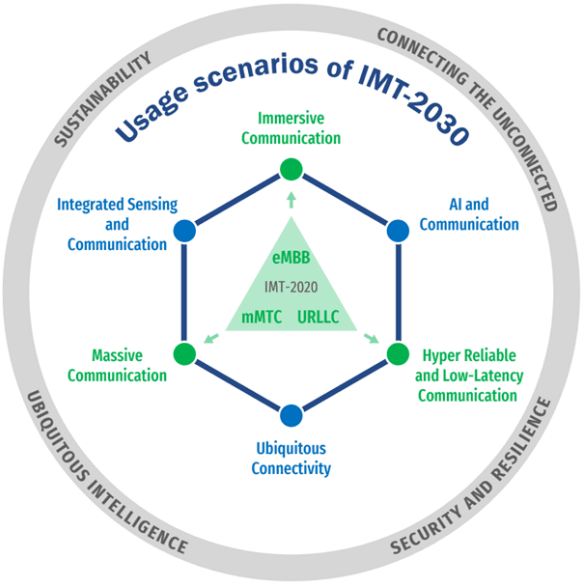


Figure 1.1: 6G Service Categories

The URLLC advances to Hyper-Reliable Low-Latency Communications (HRLLC), tightening the latency and reliability requirements to support even more demanding applications such as autonomous systems and digital twins. Three new service categories are integrated into 6G, namely the Ubiquitous connectivity (UC), Integrated AI and communication (IAC) and Integrated sensing and communication (ISAC). UC is responsible for offering seamless and continuous network coverage across all environments, including remote, aerial, and non-terrestrial domains. This occurs by integrating terrestrial, satellite, aerial, and Wi-Fi network segments into a unified infrastructure. The IAC is a novel 6G service category that embeds AI into the communication framework itself, enabling the network to autonomously adapt to the dynamic network conditions. The last service category is the ISAC, which unifies the sensing and communication functionalities within a shared spectrum and hardware infrastructure. Concurrently, 6G defines four different design directions that each service must follow, namely sustainability, connecting the unconnected, security and resilience, and ubiquitous intelligence, as depicted in Figure 1.1.

At the time of writing this thesis, no official architectural specification for 6G has been released by the 3GPP. 3GPP studies that will define the 6G radio and system architecture are expected to be completed by the first quarter of 2027, since Release 21 marks the official start of the standardized 6G networks. Nevertheless, according to the existing literature, 6G is anticipated to represent an evolution of the 5G O-RAN as mentioned in [10]. The O-RAN framework is the best candidate toward the 6G vision since it offers a programmable RIC-

driven control plane and open interfaces. However, in order to transition towards the 6G networks, some fundamental open challenges in the dynamic management and operation of disaggregated RAN infrastructures must be addressed.

This thesis contributes towards the following 6G service categories and design directions, collectively advancing the vision of 6G network management:

- **UC**: bridging indoor coverage gaps through the integration of non-3GPP WiFi access into the disaggregated RAN stack.
- **IMC**: meeting the high level bandwidth demands of B5G applications by extending the RAN's access capacity and elastically scaling its processing resources.
- **HRLLC**: ensuring low latency through dynamic and placement-aware scaling of the RAN components.
- **Sustainability**: minimizing the energy footprint of disaggregated RAN deployments through energy-aware orchestration and AI-driven power consumption observability.
- **IAC**: embedding AI natively into the RAN control and management plane as the core enabling mechanism across all of the above directions.

1.2 Open Research Challenges in 6G RAN Control and Management

To transition towards the 6G era, a fundamental rethinking of how the RAN is designed, managed, and operated is required. Although the O-RAN framework provides the architectural foundation for this transition, several open research challenges remain to be addressed before the 6G vision can be realized. The remainder of this section identifies these challenges and frames them as five interconnected research questions, each building on the previous to progressively construct the vision of an intelligent, scalable, and sustainable RAN.

6G aims to support **UC** by enabling seamless connectivity across the different wireless environments. However, real-world 5G deployments operate exclusively on licensed bands, resulting in severe indoor coverage degradation [20], thereby directly undermining this objective. One direction to bridge this gap is the integration of non-3GPP technologies, which

already dominate specific deployment scenarios, such as indoor and non-terrestrial environments, and can operate on either unlicensed or licensed spectrum. This raises the first research question:

RQ1: How can non-3GPP access technologies be integrated directly into the disaggregated RAN stack to enable unified capacity extension and service differentiation at the access layer?

However, integrating non-3GPP technologies into the disaggregated RAN stack is only the first step. The resulting multi-RAT environment introduces a new control challenge: how to distribute traffic across the heterogeneous RATs. Static traffic distribution policies cannot fully exploit the potential of a multi-RAT environment because they cannot anticipate congestion or autonomously adapt to the dynamic, heterogeneous nature of the multiple wireless channels. 6G's **IAC** service category envisions AI as a native component of the network control plane. Leveraging this capability, AI-driven traffic steering can proactively steer load across the available RATs based on predicted demand patterns. This raises the second research question:

RQ2: Given a heterogeneous multi-RAT RAN resulting from the integration of non-3GPP access technologies, how can AI-driven control mechanisms within the O-RAN architecture enable proactive and intelligent traffic steering across the available RATs?

Extending the RAN's capacity by integrating heterogeneous RATs and intelligent traffic steering increases the aggregated traffic volume the RAN processes. However, this increased radio access capacity shifts the performance bottleneck deeper into the disaggregated RAN architecture. The User Plane (UP) components of the RAN are responsible for processing and forwarding the aggregated data traffic and become the critical constraint under peak demand. In current disaggregated deployments, RAN's UP resources are statically provisioned, leading either to congestion when demand exceeds the provisioned capacity or to unnecessary over-provisioning during periods of low utilization. The cloud-native nature of the disaggregated RAN, however, opens a new opportunity to address this bottleneck: the dynamic scaling of RAN's User Plane component, enabling the network to elastically adapt its capacity in response to traffic fluctuations. Such scaling should be autonomous and AI-based to sustain the high throughput requirements as imposed by the **IMC** and the **IAC** service category of 6G. This raises the third research question:

RQ3: As intelligent traffic steering increases the aggregated traffic processed by the

RAN, how can cloud-native User Plane functions be autonomously scaled to sustain service performance under dynamic and unpredictable traffic conditions?

While dynamic scaling addresses the performance bottleneck of the RAN's UP, it simultaneously introduces a new dimension of complexity. Scaling decisions affect both the end-to-end latency experienced by the users, since latency depends on the physical placement of the RAN components within the infrastructure, and the energy footprint of the RAN, since deploying more instances requires the system to consume more power. Accordingly, 6G's scaling strategies have to jointly optimize performance, latency and energy efficiency, as it is required by the **sustainability** design direction and the **HRLLC** service category. This raises the fourth research question:

RQ4: Once dynamic User Plane scaling is established, how can different scaling strategies be designed to jointly optimize performance, latency, and energy efficiency in disaggregated O-RAN deployments?

The energy-aware scaling strategies addressed by RQ4 rely on a critical assumption: that the power consumption of each RAN component can be measured. However, in real-world deployments, measuring the power consumption of the RAN's lower-layer components, namely the DU and RU, through dedicated hardware power monitors is neither scalable nor practical. The O-RAN framework offers KPI telemetry, exposing different network metrics of the RAN components, so a fundamental question arises whether these metrics are sufficient to be used in order to train an AI model to predict the actual power consumption of the physical RAN components. Being able to predict the power consumption of the DU and RU based on the reported Key Performance Indicators (KPIs), enables autonomous power monitoring, anomaly detection, even measurements calibration without relying on dedicated external measurement infrastructure, enhancing the **sustainability** and **IAC** design directions of 6G. This raises the fifth research question:

RQ5: Given that energy-aware scaling relies on knowing the power consumption of RAN components, to what extent can network KPIs predict this consumption for the DU and RU, and how can such predictions support autonomous anomaly detection in online O-RAN deployments?

These five research questions define the research landscape of this thesis, a path from the physical extension of the RAN through heterogeneous access integration, to the intelligent and autonomous management of its resources, all the way to the energy-aware observability

of the underlying infrastructure.

1.3 Thesis Contributions to 6G Network Management

The different research questions that defined in Section 1.2 are addressed in this thesis through five distinct but interconnected contributions, enhancing the vision of intelligent, flexible, and sustainable RAN management of the 6G networks.

1.3.1 Enabling Multi-RAT Access via WiFi DU Integration in 5G RAN

Addressing RQ1, Chapter 3 proposes a novel architecture that integrates a WiFi DU into a real-world disaggregated 5G C-RAN stack. Unlike existing 3GPP solutions (N3IWF and TNGF), which operate exclusively at the CN level, this contribution implements RAN-level convergence of heterogeneous RATs. Building on this, the contribution additionally aligns the WiFi technology with the dynamic and flexible capabilities of 5G, such as network slicing. A User-Based Slicing scheme is integrated into the WiFi mechanism using Software-Defined Networking (SDN), combined with the proper exploitation of the Type of Service (ToS) field of IP packets to prioritize traffic and manage network resources. The proposed system is experimentally validated on a real-world testbed, demonstrating that improved network capacity and service differentiation are feasible in a multi-RAT 5G deployment without requiring additional licensed spectrum.

1.3.2 Proactive, ML-Driven Traffic Steering in Multi-RAT O-RAN Systems

Addressing RQ2, Chapter 4 contribution elevates the RAN-level WiFi DU integration introduced in Section 1.3.1 from a C-RAN architecture into a fully O-RAN compliant one, exploiting the O-RAN's control plane to take packet-level traffic steering decisions across the heterogeneous RATs. An xApp was developed, utilizing the Near-RT RIC to make traffic steering decisions based on UE demand forecasting using a Long Short-Term Memory (LSTM) model. The xApp applies a non-prioritizing policy that offloads or splits traffic to the WiFi DU when the 5G capacity is exceeded. The system is experimentally validated on a real-world testbed, demonstrating that the proposed mechanism dynamically adapts to the

traffic fluctuations, ensuring fair resource distribution among UEs, and maintains throughput performance without degradation even under high contention conditions.

1.3.3 On the feasibility of Dynamic Scaling of the 5G O-RAN User Plane

Addressing RQ3, Chapter 5 examines the scalability challenges of the CU-UP component, which is identified as a bottleneck in the 5G disaggregated RAN, particularly under high traffic loads. This thesis proposes a proactive horizontal scaling mechanism for the CU-UP that utilizes ML to forecast traffic demand. This involves using an xApp on the Near-RT RIC that integrates the pre-trained LSTM model to predict traffic demand, enabling proactive scaling of CU-UP instances ahead of peak periods to maintain users' requested Quality of Service (QoS). The proposed mechanism was implemented and evaluated using real-world testbeds under dynamic conditions, demonstrating its effectiveness in practical environments. Our findings emphasize the importance of integrating ML techniques to accurately forecast network loads and adapting scaling mechanisms in 5G systems to reduce packet loss and unnecessary scaling overhead, while ensuring smoother operation during high-demand periods.

1.3.4 Energy-Efficient Proactive CU-UP Scaling in 5G O-RAN: Vertical and Latency-Aware Horizontal Strategies

Addressing RQ4, Chapter 6 investigates how traffic-aware scaling of the CU-UP component can be used to reduce the operational power consumption of a fully disaggregated 5G O-RAN network. Two AI-driven xApps are deployed on the Near-RT RIC that implement proactive CU-UP scaling based on ML traffic forecasting. The first xApp implements a vertical scaling strategy by dynamically adjusting the CPU resources allocated to a CU-UP instance to minimize the 5G O-RAN power consumption, while the second performs latency-aware horizontal scaling by deploying and placing multiple CU-UP instances across the infrastructure to jointly optimize energy consumption and end-to-end latency. The proposed approaches are evaluated on a 5G O-RAN testbed using traffic traces from a real-world dataset. Experimental results highlight the applicability of scaling for the CU-UP component. Results demonstrate that vertical scaling maximizes energy efficiency, achieving up to 23% energy savings during underutilized regimes, 11% during sub-peak periods, and 8% over the full execution compared to a baseline static provisioning. In contrast, horizontal scaling presents

a complementary trade-off, achieving up to 16% savings during underutilized regimes, 7% during non-peak periods, and 4.7% for the whole execution, while significantly reducing end-to-end latency by optimally placing CU-UP instances. The results highlight that ML-driven CU-UP scaling is a practical and effective mechanism to reduce power consumption, improve energy efficiency, and enhance the performance of O-RAN 5G networks.

1.3.5 KPI-Driven Power Consumption Prediction and Anomaly Detection in O-RAN Networks

Addressing RQ5, Chapter 7 investigates the feasibility of predicting the power consumption of O-RAN components using only 3GPP-compliant network Key Performance Indicators (KPIs). We present an AI-driven framework that performs power prediction, anomaly detection, and automatic calibration of hardware measurements. The proposed approach utilizes KPI-driven ML models, trained on data collected from a real 5G O-RAN deployment. Hybrid XGBoost and CatBoost models are used to estimate the power consumption of the O-DU and O-RU nodes. Residual analysis between predicted and measured power enables the detection of sensor anomalies and measurement drift, followed by automatic calibration. Experimental evaluation of our proposed framework on a real-world testbed demonstrates accurate power prediction, with MAPE values of 1.02% for the RU and 1.22% for the DU, and a 100% detection rate for manually injected anomalies.

These contributions collectively form a research agenda that advances the vision of intelligent, autonomous, and energy-aware RAN management for 6G networks. The first two contributions extend RAN capacity and control by integrating heterogeneous RATs and AI-driven traffic steering, addressing the **UC** and **IAC** service categories of 6G. The third and fourth contributions focus on the dynamic scaling of RAN's UP components and its energy-efficient operation, targeting the performance and latency requirements imposed by the **IMC** and **HRLLC** services, while advancing the **sustainability** design direction of the 6G networks. The fifth contribution completes the picture by enabling power prediction and autonomous anomaly detection for RAN components, providing the observability layer required for trustworthy, energy-aware RAN management. Each contribution builds upon the previous one, collectively providing a solid, experimentally validated step toward the autonomous and sustainable RAN architectures envisioned by 6G networks.

1.4 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 - Background presents the foundational concepts, architectures, and technologies that are required to support the contributions of this thesis. It introduces the evolution of cellular networks from monolithic architectures to cloud-native 5G RAN, covering the O-RAN framework and its RIC-driven programmability, and defines the AI-native 6G workflow within an O-RAN system. Finally, the non-3GPP access technologies, software tools, the real-world testbed infrastructures, and the dataset used for the experimental validation of the proposed mechanisms are described.

Chapter 3 - Enabling Multi-RAT Access via WiFi DU Integration in 5G RAN addresses the RQ1 by presenting the design and the implementation of a novel architecture that integrates a WiFi DU into a real-world disaggregated 5G C-RAN stack, enabling the RAN-level convergence of heterogeneous RATs and the user-based network slicing across the heterogeneous RATs.

Chapter 4 - Proactive, ML-Driven Traffic Steering in Multi-RAT O-RAN Systems addresses the RQ2 by extending the multi-RAT architecture of Chapter 3 into a fully O-RAN-compliant framework, introducing an xApp-based traffic steering mechanism that utilizes an LSTM-based demand forecasting to proactively distribute the traffic across the available heterogeneous multi-RAT DUs.

Chapter 5 - On the Feasibility of Dynamic Scaling of the 5G O-RAN User Plane addresses the RQ3 by presenting the design and the implementation of a proactive horizontal scaling mechanism for the CU-UP component, leveraging the ML-based traffic forecasting within the Near-RT RIC control loop to autonomously adapt the RAN UP capacity under the dynamic and unpredictable traffic conditions.

Chapter 6 - Energy-Efficient Proactive CU-UP Scaling in 5G O-RAN: Vertical and Latency-Aware Horizontal Strategies addresses the RQ4 by investigating the trade-offs between the vertical and horizontal CU-UP scaling strategies in terms of the energy efficiency and the latency compliance, proposing two AI-driven xApps that jointly optimize the power consumption and the service performance of the disaggregated 5G O-RAN deployments.

Chapter 7 - KPI-Driven Power Consumption Prediction and Anomaly Detection in O-RAN Networks addresses the RQ5 by presenting an AI-driven framework that leverages the O-RAN KPI telemetry to predict the power consumption of the physical O-DU and O-RU

components, enabling autonomous anomaly detection and the energy measurement calibration in the online disaggregated RAN deployments.

Chapter 8 - Conclusion summarizes the contributions of this thesis, discusses the experimental findings and their implications for the transition toward the 6G networks, and outlines the directions for future work.

Chapter 2

Background

This chapter presents the background concepts, architectures, and technologies that form the foundation of this thesis required to support the design and the evaluation of the dynamic control mechanisms targeting the B5G & 6G networks. It introduces the principles of the monolithic 5G network architecture, and then progresses to disaggregated and cloud-native implementations, O-RAN-based architectures, and the O-RAN AI/ML workflow that underpins the intelligent mechanisms developed in this thesis. Finally, it covers non-3GPP access technologies, software tools, datasets, and testbed infrastructure used to experimentally validate the proposed mechanisms.

2.1 From analog to digital Cellular networks

1G networks were designed to support analog transmission and basic voice services with limited reliability and performance. The introduction of the 2G cellular networks marked the shift to cellular communication, enabling additional services such as SMS. The 3G era extended the cellular networks' capabilities, moving to a packet-switched architecture by separating the cellular network into CN and RAN components, enabling mobile internet access and multimedia services with limited data rates and high latency. 3G's descendant namely the 4G represents a fundamental architectural shift, introducing the first fully packet-switched, all-IP mobile network. Although the splitting of the cellular networks into the CN and the RAN was introduced in the 3G era, it is in 4G where this separation becomes fully standardized within an all-IP architecture. The CN, named as the Evolved Packet Core (EPC), is largely centralized and designed to manage mobility, session control, and data routing,

separated into the following components:

- Mobility Management Entity (MME): responsible for control-plane signaling and mobility management
- Packet Data Network Gateway (P-GW): provides connectivity to external IP networks and enforcing policy and charging rules
- Serving Gateway (S-GW): manages user-plane data during handovers
- Home Subscriber Server (HSS): maintains user subscription and authentication information

Despite these advancements, 4G networks faced different structural limitations. The centralized nature of the EPC and the monolithic implementations of RAN suffered from scalability issues, static, overprovisioned resource management, and a lack of support for heterogeneous service requirements, such as URLLC or MMTC. The aforementioned limitations highlighted the need to design more flexible, scalable, and service-oriented cellular networks, paving the way for the design and standardization of 5G.

2.2 5G Architecture

Real-world 5G deployments have two primary architectural designs, namely the Non-Standalone (NSA) and the Standalone (SA), as shown in Figure 2.1.

Non-Standalone (NSA)

In the NSA configuration, 5G is introduced only in the RAN, with a CN derived from 4G that maintains essential functionalities while incorporating some 5G features. In an NSA setup, both Base Stations (BSs)—the 4G BS, namely the eNodeBs (eNBs), and the 5G BS, namely the next-generation NodeBs (gNBs)—are connected to the same CN and can communicate directly with each other via the X2 interface. Due to these architecture principles, the NSA deployments are able to achieve greater improvements in data rates, capacity and latency compared to the 4G, primarily due to the enhancements at the RAN level. However, since the deployed CN in a 5G NSA architecture remains monolithic and does not follow a service-based architecture, critical capabilities such as network slicing and the flexibility offered by cloud-native systems are not supported.

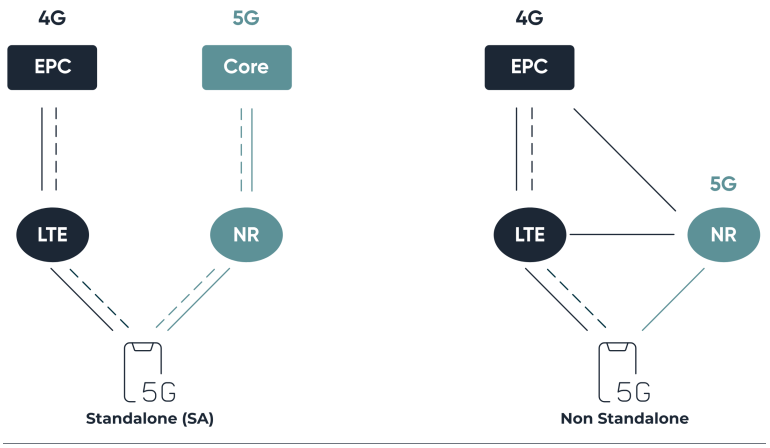


Figure 2.1: 5G NSA vs SA architectures

Standalone (SA)

On the contrary, the SA configuration aims to address the growing demand for high data throughput and ultra-low-latency services by adopting a more flexible, distributed architecture. The SA architecture leverages Network Function Virtualization (NFV) and introduces a Service-Based Architecture (SBA) across both the CN, aiming to address the limitations of NSA monolithic designs. As a result, the 5G CN has evolved into a cloud-native system by performing decomposition of the monolithic 4G CN into modular components.

Specifically, the 5G CN is decomposed into the following key entities:

- Authentication Server Function (AUSF): Responsible for user authentication, replacing the authentication functionalities of the HSS in 4G
- Unified Data Management (UDM): Manages subscriber-related data and policies, consolidating roles previously handled by the HSS
- Unified Data Repository (UDR): Serves as the centralized data storage system for network functions, corresponding to the database component of the HSS in LTE
- Network Repository Function (NRF): Enables dynamic discovery and registration of network functions, a capability not present in LTE, where functions are statically configured
- Access and Mobility Management Function (AMF): Handles access control and mobility management, effectively replacing the MME in LTE

- Session Management Function (SMF): Manages session establishment and traffic routing, assuming responsibilities previously divided between the SGW and the PGW in LTE
- User Plane Function (UPF): Implements data forwarding and packet processing in the user plane, inheriting the data-plane functionalities of the SGW and PGW.

SA architecture has a defining feature compared to the NSA and previous generations of CN architectures: the separation of network functions into Control Plane (CP) and User Plane (UP) entities, referred to in the literature as Control and User Plane Separation (CUPS). The UP components are responsible only for forwarding data traffic, while the CP entities are handling signaling and control operations. The entire architecture is shown in Figure 2.2. The existence of CUPS is also inherited in the 5G NR protocol stack, where control and user plane protocols are separated, resulting in a functional decomposition of the 5G NR entities.

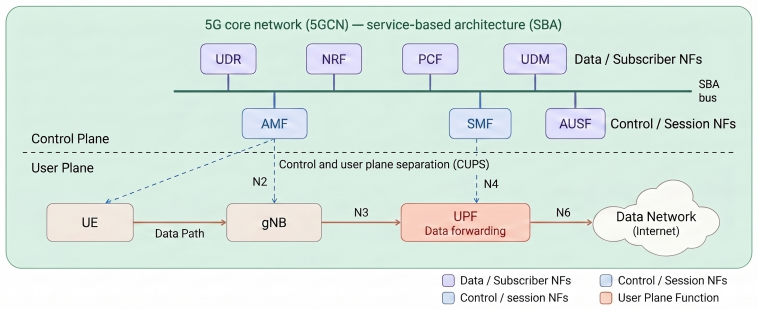


Figure 2.2: 5G SA CN

2.2.1 5G NR Protocol Stack Overview

At this point, the protocol stack of the 5G NR CP and UP will be presented. Our description will follow a top-down approach, beginning from the higher-layer protocols all the way through the lower layers that are shared between both planes

The protocols that implement the control plane protocol stack are the following:

- NAS (Non-Access Stratum): Located above the access network, NAS enables direct communication between the UE and the 5G Core (specifically the AMF) and supports procedures such as authentication, registration, and session management.
- RRC (Radio Resource Control): Responsible for connection establishment and release, mobility management, handover procedures, and configuration of security parameters.

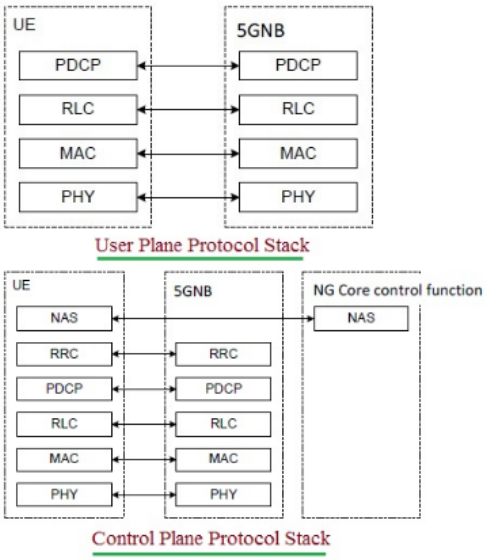


Figure 2.3: Overview of 5G NR Protocol Stack CP & UP

- PDCP (Control Plane use): Provides encryption and integrity protection for control signaling (e.g., RRC messages), ensuring secure transmission.

The UP specific layer is:

- PDCP (Packet Data Convergence Protocol): In the UP, PDCP handles data encryption, header compression, and in-sequence delivery of user traffic.

The lower layers, which are common to both control and user planes, are:

- RLC (Radio Link Control): Ensures segmentation and reassembly of data, supports ARQ-based retransmissions, and maintains in-sequence delivery.
- MAC (Medium Access Control): Performs scheduling, multiplexing, prioritization, and manages HARQ operations.
- PHY (Physical Layer): Handles the actual transmission over the air interface, including modulation, coding, MIMO processing, and beamforming.

The CP and UP plane protocols are illustrated in Figure 2.3.

2.2.2 Disaggregated and Cloud-Native Network Architectures

Historically, the RAN has followed an architecture in which its functionality is decomposed into three main components: the Baseband Unit (BBU), Remote Radio Unit (RRU),

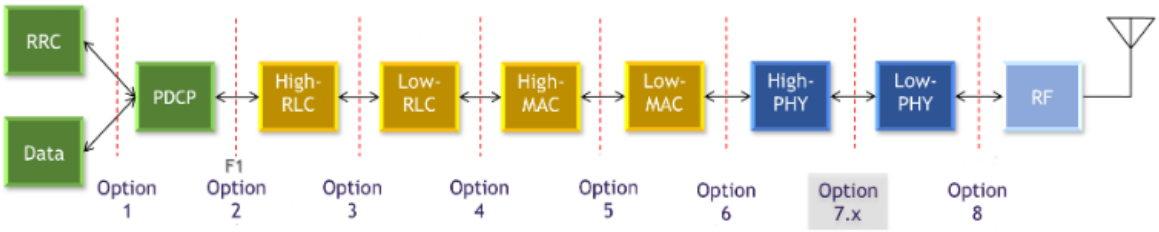


Figure 2.4: RAN Functional Splits

and the antenna. The BBU component is responsible for managing the entire BS and performs the digital signal processing, security implementations, as well as different baseband functions such as resource optimization and error detection. The BBU connects to the CN via the backhaul interface and to the RRU via the fronthaul. The RRU component is responsible for interacting with the antenna and the BBU, converting RF signals into data signals and vice versa.

According to this architectural split of the BS, different RAN deployments have been introduced:

D-RAN (Distributed or traditional RAN)

The D-RAN architecture was the dominant deployment approach adopted from the 3G & 4G era. In this approach, the BBU and the RRU are collocated at the cell site, and multiple D-RANs are connected to the CN via the backhaul. This monolithic approach suffers from reduced flexibility, resulting in poor scalability, static resource allocation, and high operational complexity.

C-RAN (Cloud/Centralized Radio Access Network)

To address these limitations, in the C-RAN architecture, the BBU is decoupled from the RRU by placing it at the hub site while the RRU remains at the cell site. This placement enables a pool of BBUs to serve multiple RRUs simultaneously, improving RAN resource utilization and enabling coordinated management across multiple cells. Building on this, 5G further disaggregates the RAN architecture by decomposing the BBU's functionality into two logical entities: the Central Unit (CU) and the Distributed Unit (DU). To support this decomposition, 3GPP introduced the midhaul interface, which is responsible for communication

between the CU and DU entities. In order to implement the aforementioned functional split of the BBU into the CU and the DU, different split points were proposed based on the 5G NR protocol stack, as depicted in Figure 2.4. Splits that occur at the lower layer, such as Option 7.2 and Option 8, place the split closer to the PHY layer, requiring very high-bandwidth, low-latency connections at the fronthaul. By choosing higher-layer splits, such as Option 2, which splits the CU-DU at the PDCP/RLC boundary, they avoid the strict requirements of lower-layer splits while maintaining the flexibility and scalability of C-RAN disaggregation. This split implements disaggregation of the BBU into the CU, responsible for higher-layer protocols (RRC, SDAP, parts of PDCP) and the DU that hosts the lower-layer protocols (RLC, MAC, PHY). The CU, following the CUPS separation adopted in the 5GCN, can even be split further into the CU-Control Plane (CU-CP) and the CU-User Plane (CU-UP). While CU-CP is responsible for the control plane signaling, the CU-UP handles the Service Data Adaptation Protocol (SDAP) marking and several Packet Data Convergence Protocol (PDCP) functions, including ciphering, header compression, packet routing and switching, as well as traffic management sourced from multiple connected DUs. In this evolved C-RAN architecture, the RRU has been redefined as the Radio Unit (RU), which remains at the cell site and is responsible for the actual over-the-air transmission and reception, implementing the lower-layer PHY and RF functions. Although C-RAN integrates elasticity into RAN deployments, it still faces important limitations, such as the need for purpose-built hardware, which limits the ability to independently scale network functions on demand.

V-RAN (Virtualized RAN)

Virtualized RAN, namely the V-RAN, overcomes the aforementioned limitations of C-RAN by extending the C-RAN approach and integrating NFV principles into disaggregated RAN components. Instead of utilizing purpose-built hardware to deploy the CU and DU components, V-RAN implements them as VNFs on commercial off-the-shelf (COTS) servers, transforming the 5G RAN from bare-metal to cloud-native deployments and enabling elasticity, resource pooling, and on-demand scaling. V-RAN lays the foundation for flexible cloud-based RAN deployments; however, it still lacks a control plane and management layer to enable intelligent, dynamic optimization of the 5G RAN.

O-RAN

Although V-RAN has resolved many challenges of traditional monolithic and C-RAN deployments, it still lacks a standardized mechanism for radio resource management and real-time adaptation. To address these limitations, the O-RAN architecture was introduced, extending the V-RAN by adding three new components: the Near-Real-Time RIC (Near-RT RIC), the Non-Real-Time RIC (Non-RT RIC), and the Service Management and Orchestration (SMO) framework, aiming to enable the RAN reconfigurability built upon two core principles: openness and intelligence.

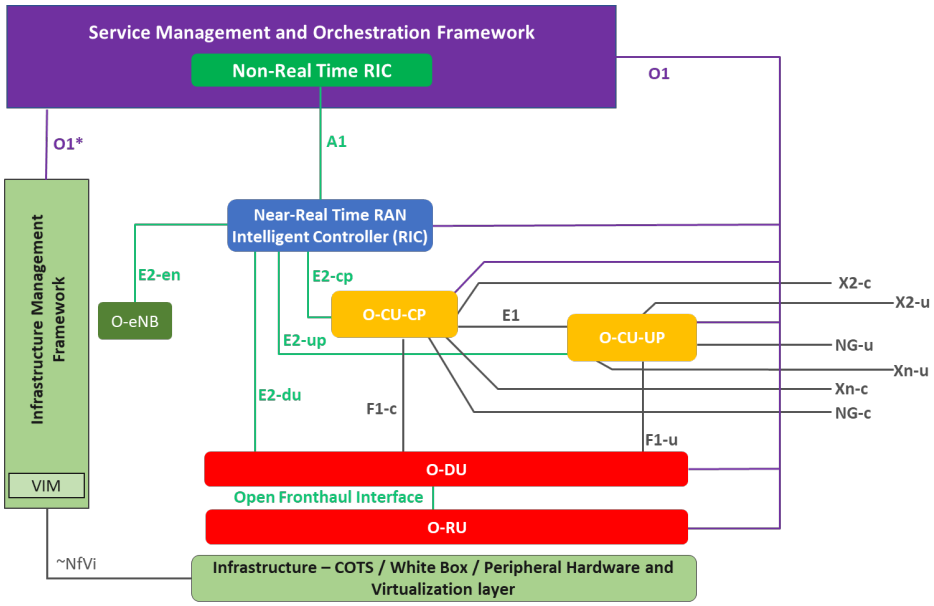


Figure 2.5: O-RAN architecture

Openness: O-RAN standardized three additional open interfaces, namely the E2, A1, and O1, in order to enable seamless interaction and communication among the multi-vendor disaggregated RAN components. Specifically, the E2 interface is responsible for connecting the near-RT RIC, to the O-CU, O-DU, where the A1 interface allows the Non-RT RIC to provide policy configurations to the near-RT RIC, and the O1 interface connects each O-RAN component with the SMO to manage all network functions, as illustrated in Figure 2.5.

Intelligence: O-RAN integrates the intelligence in the RAN via the integration of the two RICs. The Near-RT RIC is responsible for short-term optimizations such as handover management, interference control, and traffic steering, operating within a control loop of 10 ms to 1 s through modular software applications called xApps. The Non-RT RIC operates on timescales greater than 1s, handling higher-level tasks such as policy management and is

deployed as rApps.

2.3 The O-RAN AI/ML Workflow

As discussed in Chapter 1, the O-RAN architecture of 5G creates the foundational control plane for realizing the AI-native 6G vision defined by the ITU-R IMT-2030 framework [19] with the IAC service, where ML-driven proactive management becomes an architectural necessity rather than an engineering preference.

The O-RAN architecture, introduced in Section 2.2.2, provides a ready-to-deploy framework for implementing real-world cellular networks with AI-native applications. For an architectural embedded AI system, there are two required layers in order to define the end-to-end ML workflow: Model training and model inference.

Model training is performed at the SMO framework, where historical KPI data are collected via the O1 interface and the different ML models can be trained offline, without real-time constraints.

Model inference in O-RAN can happen at two levels depending on the latency requirements of the use case. For Non-RT applications, the ML models are fetched from the SMO at the Non-RT RIC and executed directly as rApps at the Non-RT RIC, interacting with different RAN components through the O1 interface. On the contrary, for the near-RT applications, the Non-RT RIC is responsible for receiving the pretrained model from the SMO and delivering it to the Near-RT RIC via the A1 interface, where the xApps receive the model and perform the inference.

All the AI/ML-driven mechanisms proposed in this thesis follow the aforementioned workflow, being a proof of concept for the AI-native network management paradigm envisioned by 6G, validated through real-world deployments on state-of-the-art 5G O-RAN testbeds.

2.4 Non-3GPP Access Technologies

As introduced in Chapter 1, to achieve the 6G vision of UC, the integration of heterogeneous RATs is required. This section presents the non-3GPP technologies relevant to this thesis, with a focus on WiFi as the primary non-3GPP RAT candidate.

RATs can be generally categorized into two main groups from a cellular network perspective, depending on the type of access: 3GPP and non-3GPP. The first one includes access technologies that are standardized by the 3GPP, such as the 4G LTE and 5G NR, mainly operating in licensed spectrum and supporting mobility management, QoS and other key RAN technologies. However, non-3GPP technologies have distinctive characteristics and operate outside of the 3GPP standard. The two most important non-3GPP RATs are Wi-Fi, which provides high-capacity indoor offloading, and Satellite, which offers ubiquitous global reach. In particular, WiFi (IEEE 802.11) operates in the unlicensed spectrum and offers affordable, high-bandwidth connectivity suitable for indoor and local-area deployments. On the contrary, satellite systems operate in licensed frequency bands (Ka- and Ku-bands) to provide global coverage.

The integration of non-3GPP RATs into 3GPP-based systems is motivated by their complementary characteristics. The convergence of these technologies allows operators to combine the strengths of both, enabling flexible and efficient access networks that can better address diverse service requirements and deployment scenarios. Conducting real-world research and experiments using satellite communications is particularly challenging due to the lack of open, accessible testbeds for this technology and the extremely high infrastructure costs of independent deployments. On the contrary, WiFi offers a low-cost and easily accessible non-3GPP access technology, making it suitable for real-world experimentation. For these reasons, in this thesis WiFi was selected as the reference non-3GPP access technology for the integration of non-3GPP components into 5G.

2.4.1 IEEE 802.11 - WiFi

WiFi follows a star topology, in which all wireless devices, known as stations (STAs), communicate through a central node called the Access Point (AP). The AP is the bridge between the wireless client and the chosen wired network. This topology is depicted in Figure 2.6.

The WiFi protocol stack, based on IEEE 802.11, follows the OSI model and mainly focuses on the PHY and MAC layers. The PHY layer handles modulation, coding, and transmission over the radio channel using different techniques such as DSSS, OFDM, or MIMO depending on the standard. The MAC sublayer manages the medium access using the CSMA/CA mechanism, ensuring collision avoidance and reliability through acknowledgments and re-

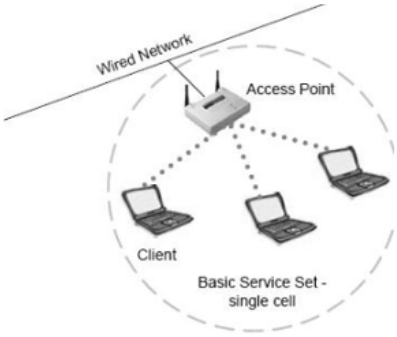


Figure 2.6: WLAN architecture

transmissions. Higher layers, such as the network and transport layers, remain protocol-agnostic, allowing seamless integration with TCP/IP networks.

In this thesis, the IEEE 802.11ac, also known as WiFi 5, was used as the non-3GPP RAT. It was introduced in 2013 and operates at the 5 GHz frequency band. WiFi 5 supports wider channels of 80 and 160 MHz, higher-order modulation (up to 256-QAM), and Multi-User MIMO (MU-MIMO) in the DL. It employs CSMA/CA for medium access control and reaches a maximum theoretical data rate of approximately 7 Gbps making it the best candidate for high-bandwidth applications and real-time communication.

Despite these advantages, WiFi's operation in unlicensed spectrum introduces a fundamental vulnerability: interference. Since unlicensed bands are shared with other devices and networks, WiFi performance can degrade significantly in dense or uncontrolled radio environments due to co-channel and adjacent-channel interference. However, addressing interference management is outside the scope of this thesis. The experiments conducted in Chapters 3 and 4 are performed in an RF-isolated environment, where interference effects are controlled and do not impact the validity of the results.

2.5 Deployment Tools

This section presents the software tools and platforms used in this thesis to experimentally validate the proposed mechanisms on real-world testbeds. The selected tools cover the deployment of all components of a 5G network (namely the 5GCN, O-RAN, and RICs), the WiFi components for the convergence of non-3GPP RATs to 5G, as well as tools for containerization following NFV principles and for the cloud orchestration platform.

2.5.1 5G software suites

The main 5G software platform used in this thesis is the OpenAirInterface (OAI) [21], which provides an open-source implementation of the NR protocol stacks maintained by the OpenAirInterface Software Alliance (OSA). OAI provides software-based, NFV implementations of the 5GCN and RAN, offering an environment with the support of D-RAN, C-RAN, V-RAN and O-RAN deployments on commodity off-the-shelf hardware.

Focusing more on the 5GCN part, OAI 5GCN provides implementations of all the key 5G network functions, including AMF, SMF, UPF, AUSF, UDM, UDR, and NRF. These VNFs are deployed as containerized services and constitute the control-plane and user-plane backbone of the experimental 5G deployments used in Chapters 3, 4, 5, and 6.

Regarding the RAN part, as already mentioned, different implementations following D-RAN, C-RAN, V-RAN, and O-RAN deployments are provided. OAI-RAN also includes an RF Simulator that emulates the physical radio layer using an Additive White Gaussian Noise (AWGN) channel model, enabling controlled experimentation without physical RF hardware.

For the O-RAN RICs, this thesis relies on OAI FlexRIC [22], an open-source Near-Real-Time RIC tightly integrated with OAI-RAN via the E2 interface. FlexRIC provides the run-time environment for xApps, which monitor and control RAN components on timescales ranging from 10 ms to 1 s by continuously collecting KPMs from E2 nodes via E2 Service Models such as E2SM-KPM. FlexRIC is used in Chapters 4, 5, and 6 as the platform for the traffic steering and CU-UP scaling xApps developed in this thesis.

In Chapter 7, the 5G deployment is not based on OAI but instead utilizes the Accelleran O-RAN software stack, offering CU, DU, and Near-RT RIC deployments on a fully disaggregated physical testbed with a real radio unit. In the same Chapter, Open5GS is adopted as the 5GCN, deployed as containerized services alongside the RAN components.

2.5.2 WiFi Access Point Software

For the scope of this thesis, all the WiFi components were deployed using the **ath10k** driver [23]. Ath10k is an open-source Linux kernel driver designed for the Qualcomm Atheros 802.11ac wireless chipset families. It is responsible for providing the low-level interface between the operating system and the wireless NIC, such as firmware loading, hardware initialization, and radio configuration.

For the AP deployments, we used the **hostapd** [24], an open-source user-space daemon

for Linux that implements IEEE 802.11 Access Point functionality. It is responsible for managing all AP-side operations, such as client authentication and association, beacon transmission, SSID advertisement, and other related tasks. Hostapd communicates with the kernel through the nl80211 drivers, which support a wide range of IEEE 802.11 protocols. In this thesis, it enables the deployment of software-defined WiFi access points for integrating non-3GPP access into the 5G system. According to the literature, hostapd is the de facto standard tool for software-defined AP deployments on COTS hardware.

2.5.3 Containers

Containerization is the technology that enables the NFV-based deployment adopted from the 5G & B5G networks. A container is a lightweight, isolated, and executable environment that packages an application with the required environment, including all its dependencies, libraries, and configuration. In this thesis, containers are used to deploy network functions, including the CU, DU, and 5G CN components. Containers' lightweight environment is underlined via the sharing of the host operating system kernel and relies on Linux namespaces and control groups (cgroups) for process isolation and resource control, enabling extremely low startup times, something crucial for the dynamic scaling mechanisms presented in Chapters 5 and 6. In this work, the container implementation relies on Docker [25], a widely adopted platform that provides the tooling for building, distributing, and running containerized applications.

2.5.4 Kubernetes

Kubernetes (k8s) [26] is an open-source container orchestration platform that automates the deployment of containerized workloads across cluster nodes, enabling cloud-based application deployment. In this thesis, K8s is used in order to provide the cloud-based environment needed to deploy the different CNFs following the V-RAN and O-RAN deployments. A k8s cluster is composed of a control plane node, responsible for cluster management, scheduling, and orchestration, and a set of worker nodes that execute containerized workloads. The fundamental unit in Kubernetes is a Pod, which contains one or more containers (e.g., CU, DU) that share the same network namespace. Deployments are a higher-level abstraction responsible for maintaining and managing the lifecycle of the deployed pods, while services

provide network endpoints for Pods (e.g., between CU, DU, and 5G CN) regardless of their underlying placement across the cluster.

2.6 Testbeds

This section presents the various testbed infrastructures used in this thesis to experimentally validate the proposed mechanisms. The use of real-world testbeds is a key aspect of this thesis, going beyond simulation-based approaches and demonstrating the practical applicability of these approaches in real-world 5G networks.

2.6.1 SLICES Research Infrastructure (SLICES RI)

SLICES-RI [27] is a pan-European research infrastructure, designed to support large-scale experimental research in computer science, providing technologies such as 5G/6G, IoT, cloud computing, and High Performance Computing (HPC). The SLICES-RI provides unified testbeds from 16 European Countries as a single research instrument, providing high-quality experimentation services to the research community.

SLICES-RI has multiple national nodes distributed across multiple European Countries, each one offering different experimentation capabilities. The Greek node of SLICES-RI, namely the NITOS testbed (Network Implementation Testbed using Open Source platforms), is developed and operated by the Network Implementation Testbed Laboratory (NITlab) at the University of Thessaly in Volos, Greece. NITOS is the primary testbed used in this thesis for the experimental evaluation of Chapters 3, 4, 5, and 6. NITOS is an open-access and fully configurable platform with over 100 operational wireless nodes, distributed across three different outdoor, indoor, and office deployments, all located at the University of Thessaly campus. Researchers can remotely access the infrastructure and conduct experiments with a wide range of wireless technologies, including 5G, WiFi, and SDRs, in both RF-isolated and open-air environments.

2.6.2 5GTNF

The 5G Test Network Finland (5GTNF) is a nationwide Finnish research infrastructure coordinated by the VTT Technical Research Center of Finland. The 5GTN was the world's first open 5G test network, established as a joint effort between the University of Oulu, VTT,

and multiple industry partners, and has since evolved to integrate O-RAN technology and foundational capabilities for 6G experimental research.

The 5GTN offers a 5G test network infrastructure that includes RANs utilizing various technologies and operating across multiple frequency bands. It features several CN deployments, ranging from stable carrier-grade installations to experimental configurations, as well as service development platforms tailored for diverse vertical use cases. The Oulu node of 5GTN, located at VTT Technical Research Center of Finland, serves as the experimental site for this thesis. This node provides a fully disaggregated O-RAN deployment using commercial off-the-shelf hardware, real radio equipment, and per-component energy instrumentation, which makes it particularly well-suited for the energy-aware experiments described in Chapter 7.

2.7 Dataset

The experiments in this thesis rely on a publicly available 5G dataset, selected because RAN-side measurements from commercial operators are not openly accessible, as operator-managed infrastructure metrics are not publicly released for privacy reasons. Consequently, the datasets used are collected from the UE side, which, although limited to the device perspective, provides a realistic picture of the network conditions experienced by a single user, enabling experimentation based on real-world 5G traces rather than controlled testbed environments.

The dataset utilized in the majority of the works of this thesis is the publicly available dataset in [28] that provides a 5G trace gathered from a major Irish mobile operator and offers two mobility patterns (car and static) and two application patterns (file download and video streaming) from the UE perspective. Among the candidate mobility and application patterns, we use the static one with a downloading application, as it yields the highest client bitrate demands with fewer fluctuations than the other patterns, making it particularly suitable for our study. As our goal is to stress-test the different 5G network components, replaying this portion of the dataset yields the highest possible component utilization. Furthermore, the static pattern dataset has more entries than the mobility dataset, which improves its suitability for training ML models.

The dataset comprises a rich set of client-side KPIs, as listed in Table 2.1.

Table 2.1: Dataset KPIs

KPI	KPI	KPI
Timestamp	Ping average (Ping_avg)	SNR
Longitude	Ping minimum (Ping_min)	RSRQ
Latitude	Ping maximum (Ping_max)	RSRP
Velocity	Ping std (Ping_std)	RSSI
Operator name	Ping loss (Ping_loss)	CQI
Cell ID	DL bitrate (DL_bitrate)	NRxRSRQ
Network mode	UL bitrate (UL_bitrate)	NRxRSRP
State (idle/downloading)		

Overall, the combination of these features provides a comprehensive view of the UE-side network conditions, capturing temporal, spatial, throughput, latency, and radio signal characteristics. Although limited to the device perspective, this level of detail is sufficient to enable meaningful analysis and experimentation, particularly for evaluating network performance and training machine learning models under realistic 5G deployment conditions.

2.8 Conclusion

In this chapter, the foundational concepts, architectures, and technologies that underpin the contributions of this thesis are presented. The 5G standalone architecture was introduced, covering the disaggregated cloud-native RAN and the O-RAN framework with its RIC-driven programmability, while the AI-native 6G vision was presented as the motivation for embedding intelligent control into the network infrastructure. The non-3GPP access technologies, the software tools, datasets, and real-world testbed infrastructures were also described as the experimental foundation of this thesis. The concepts and tools introduced in this chapter are directly referenced throughout the following chapters: the WiFi technology is the non-3GPP RAT integrated in Chapters 3 and 4, the O-RAN architecture and the Near-RT RIC form the control plane for the mechanisms designed in Chapters 4 to 7, and the NITOS and VTT testbeds serve as the validation environments for Chapters 3–6 and Chapter 7 respectively.

Chapter 3

Enabling Multi-RAT Access via WiFi DU Integration in 5G RAN

3.1 Introduction

As defined in Chapter 1, 5G was designed to simultaneously address the three fundamentally different service categories of eMBB, URLLC and mMTC. Each of these categories has distinct demands of the RAN, for instance, eMBB requires sustained high throughput, URLLC demands deterministic low latency, and mMTC calls for massive connection density at low cost. To meet all the aforementioned three services with the usage of a single, homogeneous RAN that operates exclusively on licensed spectrum is inherently limited, since the available spectral resources are finite, while the cost of high-density licensed infrastructure deployments to satisfy peak demand in high-traffic environments such as indoor venues remains prohibitive [10]. Since 6G aims to extend these requirements further through the UC and IMC service categories by targeting sub-millisecond latency, Tbps throughput, and seamless connectivity across heterogeneous environments, the limitations of a single RAT-based RAN become untenable.

The aforementioned limitations of the licensed bands are readily evident in indoor 5G deployments. The majority of real-world 5G deployments operate at mid-band frequencies and suffer from building-entry loss and wall penetration, making outdoor-to-indoor coverage strategies fundamentally insufficient to deliver consistent eMBB performance inside urban hotspots and office buildings [20]. This situation is further compounded in higher 5G frequency bands where real-world measurements have confirmed that after penetrating a con-

crete wall the 5G mmWave peformance's DL rates drop around 93% (from 600Mbps to 41.5 Mbps) as described in [29]. The integration of non-3GPP RATs into the 5G, has been identified as a key strategy to address the above limitations, by extending network capacity into the unlicensed spectrum without additional licensed bands investment [30] or suffering from the indoor service degradation. Among the different non-3GPP Access Technologies, WiFi operates in unlicensed bands and has already been deployed at massive scale globally [31], making it the best candidate to complement 5G NR deployments in dense indoor environments.

The existing 3GPP-compliant integration mechanisms for non-3GPP access technologies into 5G operate exclusively at the CN level, limiting RAN-level coordination. The Non-3GPP Interworking Function (N3IWF), introduced in 3GPP Release 15 [32], serves as a gateway between the untrusted non-3GPP access networks and the 5G CN, establishing an encrypted IPsec tunnel between the UE and the CN while operating entirely outside the RAN stack. For the trusted non-3GPP access networks, 3GPP Release 16 [33] introduced another component at the 5GCN, the Trusted Non-3GPP Gateway Function (TNGF), having almost the same functionality as the N3IWF, with the key difference that it uses a null-encrypted IPsec tunnel since the underlying WiFi link security is already provided by the operator.

An important characteristic of the 3GPP-compliant methods for integrating non-3GPP RATs into 5G is the absence of RAN integration. This limitation motivates **RQ1**: **How can non-3GPP access technologies be integrated directly into the disaggregated RAN stack to enable unified capacity extension and service differentiation at the access layer?** In this Chapter, the disaggregated 5G C-RAN architecture is extended to support the convergence of the Non-3GPP WiFi Access Technology, as a non-3GPP DU within the 5G RAN stack, managed by the CU through the F1 interface at the PDCP/RLC split boundary. In contrast to the approach followed by N3IWF and TNGF, which operate at the CN level, this chapter introduces a MAC-level integration that enables the CU to jointly schedule traffic across 5G NR and WiFi DUs. Since this approach operates at the RAN level, slicing policies can be enforced directly at the access layer, enabling per-user QoS differentiation to be applied uniformly across both 5G NR and WiFi DUs, a capability that N3IWF and TNGF are fundamentally unable to provide. The proposed system is implemented and experimentally validated on the NITOS testbed [34], demonstrating that RAN-level WiFi DU integration is not only architecturally sound but practically feasible in operational 5G networks.

The main contributions of this chapter are the following:

- To increase 5G network capacity with non-3GPP technologies, integrated into the 5G networking stack as separate Distributed Units (DUs).
- To enable slicing information distribution and enforcement for non-3GPP technologies, complying with the 5G slicing model.
- To experimentally evaluate the contributions in a testbed environment, under real-world settings.

3.2 Related Work

Already from the first generations of cellular networks (e.g 3G), WLAN integration was considered crucial. The work [35] describes and proposes the integration of WLANs at the level of architecture and protocols as a separate part of 3G systems, which communicates through a proxy server with the PLMN in the 3G core network and not in the UTRAN access network. In the 4G era, as described in Release 12, two different approaches were introduced to converge WLAN into it: the LTE-WLAN Aggregation (LWA) and the LTE-WLAN Radio Level Integration with IPsec Tunnel (LWIP). The first method utilizes the data converge Protocol (PDCP) as the aggregation layer [36], whereas the second one uses IPsec connections based on an IP-Layer 3 aggregation [37].

The N3IWF as introduced in Section 3.1 seems to be the descendant of LWIP methods of 4G era and its conversion into a function of 5G CN. N3IWF directly interacts with the AMF, SMF, and UPF to manage mobility and access of non-3GPP UE, manage the PDU sessions, and establish data plane connections using GTPU tunnels, respectively. In [38], the authors perform an integration of WiFi and satellite technology, two non-3GPP wireless communication systems, into a 5G network utilizing the N3IWF to evaluate the performance of such integration in terms of timing of UE registration and PDU session establishment and also the RTT. The results show that the overhead introduced by the N3IWF, is negligible in real-world WiFi packets having a length of 1500 bytes. The authors of [39], provide a comprehensive overview and practical insights into the convergence, authentication, and authorization procedures involved into integrating, non-3GPP access networks, such as WiFi, into 5G networks using the N3IWF component.

One of the key enabler features of 5G is network slicing, which creates multiple virtual networks that multiplex the resources atop a cellular network to provide different services such as end-to-end connectivity and others. Performing slicing into RAN, called RAN Slicing, is one of the most challenging aspects of 5G since the dynamic wireless environment conditions prevailing into the medium in addition to limited radio resources. Authors of [40] address the challenges of inter-slice interference by proposing an algorithm using heuristics to optimize RB (Resource Block) allocation to minimize inter-slice interference in multi-slice networks.

New-era WiFi protocols such as IEEE 802.11ax and IEEE 802.11be integrate slicing methods such as Multi-SS [41], creating separated networks in a network management perspective focusing on efficiently handling multiple users' demands and improving bandwidth utilization, not resource allocation as 5G does.

Integrating a WiFi access network into a 5G system implies the need to support similar features as that of slicing in our implementation. Targeting home networks, extensive research has been conducted regarding Network slicing in WiFi. The authors of the work [42] combine FlowVisor with OpenFlow networks to implement independent and isolated slices from each other to satisfy applications from different providers or with different network requirements. Apart from the available bandwidth of the channel, the slices share other essential network resources, such as the CPU of the network devices or even forwarding table entries. A different approach was followed in the paper [43], where the researchers see WiFi slicing as a resource allocation problem as a stochastic optimization challenge, using Optimization Theory to ensure minimum guaranteed bit rates and maximum queuing delays. The solution they propose is a packet routing algorithm that will manage the airtime of the medium corresponding to each slice, in order to achieve queue prioritization. One way to perform slicing into WiFi networks is to utilize a MAC-level slicing technique. Paper [44] presents a novel approach of MAC level slicing based on pausing and unpausing the MAC80211 software queues before the data is transferred to the NIC, enabling precise allocation of air-time to different virtual networks of a single physical Access Point (AP). This mechanism allows for real-time resource isolation in time based on a TDMA approach, where the medium access is divided into small slots, allowing fine-grained adjustment of the slice size during runtime. Nevertheless, none of the above works addresses how slicing can be applied when integrating non-3GPP technologies to the cell.

In this chapter, we progress beyond the current state of the art to develop, to the best of our knowledge, a first prototype integrating WLAN as a non-3GPP DU to the 5G RAN. This chapter extends slicing capabilities to the non-3GPP network, by handling the scheduling at the PDCP level of the 5G stack, and enforcing slicing decisions to the non-3GPP network. In the next sections we detail our system architecture that enables such functionality.

3.3 System Model and Architecture

This section describes the system architecture and our method to perform the slicing over WiFi DU. In our base architecture, we use as reference network stack the 5G stack, using the 3GPP Option 2 split described in Section 2.2.2, as well as in paper [45] and demonstrated in [46], in which CU includes the protocols of the PDCP layer and above, and DU includes the RLC layer and below. CU, gNB DUs, and WiFi DUs are the primary entity components of this Heterogeneous RAN 5G Disaggregated architecture. In the following of this chapter, we analyze the implementation of WiFi DU entity and how slicing is performed.

3.3.1 WiFi DU integration

For the WiFi DU, we utilized a handler developed for the 4G OAI architecture, to fulfill the communication protocol requirements of RLC and PDCP in the disaggregated RAN implementation, allowing us to follow the LWA aggregation architecture into integrating the WiFi DU on the 5G OAI architecture. This handler intervenes between the RLC and PDCP in the protocol layer stack, providing an interface for reliable communication in the control and data plane among these layers and enabling data transfer between multiple DU technologies and the CU. The WiFi-DU registration process in the network does not follow the conventional process of 3GPP-compatible DUs. Thus, the CU is registering with the WiFi-DU, unlike 3GPP DUs, which follow the opposite process and are responsible for their registration to the CU.

The WiFi stack considerably differs from the cellular network one concerning functionality. Different procedures must be followed, depending on whether data will be sent back to the CU via the F1 interface or to network UE via the WiFi NIC on the DU device. The WiFi DU also has the AP's role in the IEEE 802.11 wireless network system, and the UEs have to establish a connection with it. Regarding the UL dataflow of WiFi DUs, packets are

received from UEs using the IEEE 802.11 protocol stack. So, the received payload (IP layer and above) should be encapsulated into the respective PDCP header for the PDCP instance running on the CU. In the DL dataflow, packets must undergo the reverse process: unpack and strip off the PDCP header and deliver the payload to the driver operating over the NIC on the DU device.

On the 5G-DU side, we use the same architecture as the 5G OAI, utilizing the F1AP protocol for control messages and the GTP-U protocol for the data traffic among the CU and the DU. Our strategy is to support one multihomed UE to be connected simultaneously to two DUs, one WiFi, and one 5G DU, in order to increase the overall network capacity. Each DU is responsible for managing and maintaining its own PDCP sequence number for the packets of each UE that it processes. Implementing a distributed individual PDCP instances architecture at each DU to our system leads to increased complexity to ensure consistency in handling PDCP sequence numbers due to the need for communication among the DUs, adding synchronization overhead to correct packet ordering and delivery for the higher layers. We chose to integrate a centralized PDCP instances architecture at the CU for unified management of sequence numbers of a UE, irrespective of the DU that processes them, by aggregating the sequence number of the PDCP entity from both DUs.

3.3.2 WiFi Slicing Mechanism

The stripped packet is an IP packet and its payload. In IP protocols, the ToS field provides a quality of service (QoS) mechanism used in networking routing functionalities to prioritize the network traffic, ensuring efficient utilization of network resources to deliver low-latency, high-throughput, or highly reliable service for the datagrams. We prioritize datagrams by changing the value of the Differentiated Services Code Point (DSCP)/Type of Service (ToS) field in the IP header. To effectively support the Quality of Service (QoS), a mac80211 driver has at least four priority queues, each of which is mapped to a different Access Category depending on its services (Table 3.1). In ascending order, these Access Categories are as follows: Background (BK), Best Effort (BE), Video (VI), and Voice (VO).

As described in the IEEE 802.11 stack protocols, each of these four access categories is mapped to a User Priority (UP). UP is a value associated with a medium access control (MAC) service data unit (MSDU) specifying the quality of service class and the priority queue that each datagram belongs. The mapping between UPs, ACs, and MQs is shown in Table 3.1.

Table 3.1: Mapping of WiFi AC and MAC Queues

User Priority	Access categories (ACs)	mac80211 queue (MQs)
1	Background (BK)	3
2		
0	Best Effort (BE)	2
3		
4	Video (VI)	1
5		
6	Voice (VO)	0
7		

The prioritization among different queues is achieved through different opportunistic channel access characteristics. More particularly, different MQs use different inter-frame periods to consider the medium idle before transmitting their datagram. Therefore, MQs with higher priorities will access the channel more frequently due to shorter inter-frame times between subsequent transmission attempts. The ToS value of the IPv4 header determines the packet UP. The IPv4 ToS field maps into the UP as shown in Table 3.2.

Table 3.2: ToS-UP mapping

ToS val	User Priority
0-31	0
32-63	1
64-95	2
96-127	3
128-159	4
160-191	5
192-223	6
224-255	7

As we mentioned, the PDCP header is taken from the WiFi DU entity's DL stream, leaving only an IP packet and its payload. We achieve slicing at the WiFi level by giving different

values to the ToS field of the IP packet depending on the destination address of the IP packet due to the existence of UP queues will be assigned.

Unlike the traditional WiFi protocols, that distribute equal channel capacity among the connected stations, our proposed system implements a User-Based Slicing approach. The framework prioritizes the clients based on the ToS value that their packets will be assigned. In this system, clients whose packets are allocated to a high-priority queue will access the channel more frequently disrupting the fair distribution of channel resources leading to their prioritization. This approach mitigates the potential starvation of lower-priority clients by leveraging the driver mechanisms designed to address this imbalance.

3.4 Evaluation

Extensive experiments were performed in this part to demonstrate the viability of the proposed idea. To evaluate our implementation, we used the SLICES-RI's NITOS testbed as described in Section 2.6.1, located in the University of Thessaly, that offers highly programmable, remotely accessible technologies to users worldwide.

3.4.1 Experiment Setup

We deployed a Heterogeneous RAN 5G Disaggregated network in the NITOS testbed, an RF-isolated environment, based on a simple architecture in which four nodes are used to deploy the 5G CN, the CU, the 5G-DU and WiFi-DU in order to evaluate our proposed network architecture. For the 5G-DU, two different USRPs, B210 and N310 were utilized to perform experiments and to compare their performance in real environments. For the UEs, two additional nodes equipped with the 5G dongle Quectel RM510Q-GL and a Qualcomm Atheros QCA988x 802.11ac WiFi NIC were used. The WiFi-DU node is also equipped with a Qualcomm Atheros QCA988x WiFi NIC to support the IEEE 802.11ac [47] protocol suite. The described architecture is shown in Figure 3.1.

The 5G CN's functionality is provided from the OAI 5G core, where we chose the basic deployment of NRF, UDR, UDM, AUSF, SMF, AMF, and UPF network functions as mentioned in Section 2.2. For the 5G RAN, we utilized the OAI-RAN. Regarding the 5G-DU, we did not make any changes to the source code provided by OAI. In contrast, for the 5G-CU, we integrated the library and a PDCP aggregation mechanism to support a UE connected to

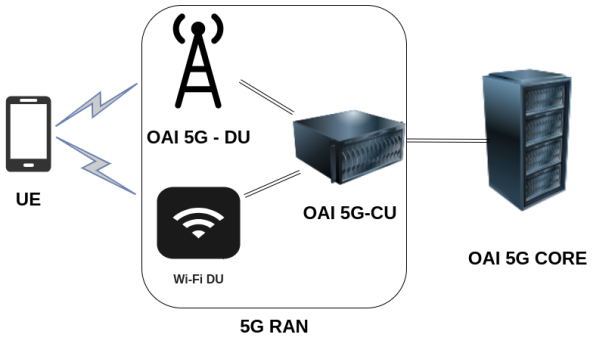


Figure 3.1: 5G Disaggregated Architecture

multiple DUs, as described in Section 3.3, and we implemented two basic selection interface policies based on Round Robin and Flooding schemes.For the WiFi-DU, we followed the same strategy as in [46]. We created a Python-based agent using the Google protocol buffers library to create a messaging exchange scheme among the handler integrated into the 5G-CU and the WiFi-DU. This agent receives the CU messages, retrieves the payload, and injects it into the AP WiFi device using a lookup table to link each UE RNTI to a WiFi NIC MAC address. The RNTI information of each UE is propagated to each WiFi DU upon registration of a new client with the 5G network, by the CU.

As proposed in the previous section, we achieve user-based slicing at the WiFi level by assigning different values to the IP packet's ToS field depending on its destination UE, according to preamble decisions. To achieve the shortest possible delay when configuring the IP packet for its prioritization, we decided not to configure this modification through appropriate libraries at the application level, but rather to pass the packets to an Open vSwitch (OvS) instance that uses programmable flows to support the appropriate modifications (ToS changes). As shown in Figure 3.2, to support two user-based slices with different priorities we connected two virtual ethernet interfaces and the WiFI NIC to an OVS. In our proposed solution, we establish a corresponding number of veths to the available ACs, to enable the assignment of each client to a distinct link that corresponds to a specific AC. Since virtual Ethernet devices are always made in pairs, two more virtual interfaces and the links required to complete the connection between these pairs were configured. One interface from the pair connects to OvS, and the other is forwarding packets to OvS from our agent. In the agent's DL dataflow, we modified the application according to the following functionality: depending on the priority we want to assign to each packet, the packet traffic is split across the two

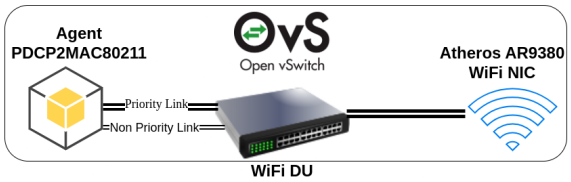


Figure 3.2: WiFi DU slicing architecture

veth interfaces. OvS is programmed to change the ToS value on packets arriving at one of the two veth interfaces and forwarding them to the WiFi interface. We change the routing based on the packet destination address, or the client, between the priority link and non-priority link every predetermined number of packets, to achieve the channel distribution in percentages.

The AP functionality of the WiFi DU was set up using the nl80211 and the ath10k_pci driver. The chosen operation wireless channel was 36, with the hardware mode set to operate in the 5 GHz band. The 802.11ac standard was enabled to leverage the benefits of High Throughput (HT) and Very High Throughput (VHT) capabilities. The VHT capabilities were fine-tuned with features such as LDPC coding having a channel width of 80 MHz and a center frequency of 5210 MHz. WiFi Multimedia (WMM) was activated for enhanced Quality of Service (QoS), playing a crucial role in our slicing mechanism. This configuration maximizes the transmission speed in the WiFi medium for our experiments.

For the 5G DU setup, we tested two USRPs, the b210 and n310, to measure their performance in real environments. We completed tests using USRPs B210 and N310, using 1x1 and 2x2 transmission schemes, to determine the best configuration that minimizes the round-trip time (RTT), increases throughput, and offers stability to the connections, the results of which will be presented in Section 3.4.2. Both USRPs operated in N78.

3.4.2 Experiments

In this subsection, we will present the results of our experimental evaluation. Firstly, we compared the 5G-DU and the WiFi DU in terms of the RTT and the throughput speed among these technologies. This comparison offers insightful information about the advantages and limitations of each technology in our testbed setting. Following this, we illustrate the results of the proposed WiFi-DU slicing scheme, in an effort to integrate RAN slicing into the WiFi-DU implementation.

Many applications that 5G aims to support are latency-sensitive applications, such as

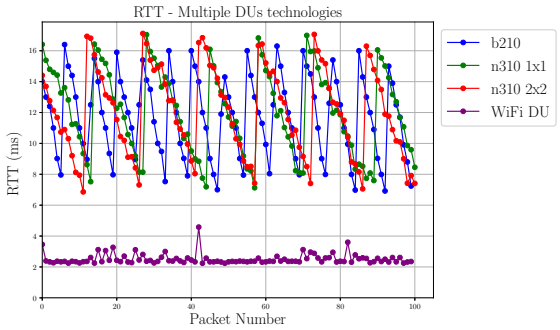


Figure 3.3: RTT of packets across multiple DUs technologies

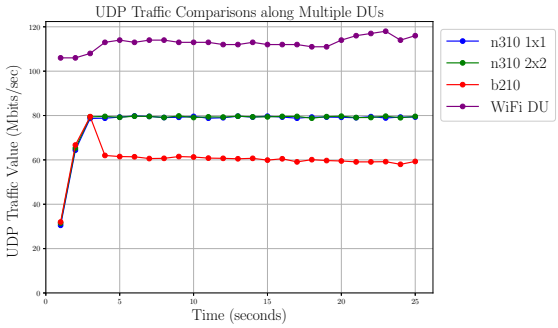


Figure 3.4: DL UDP traffic across multiple DUs technologies

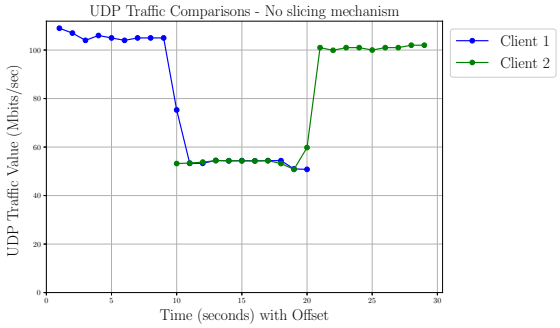


Figure 3.5: UDP transmission to two stations without slicing



Figure 3.6: UDP transmission to two stations slicing supported

virtual reality (VR), augmented reality (AR), autonomous driving, and industrial automation, highlighting as one of the most fundamental metrics to evaluate the importance of a WiFi DU integration into 5G architecture the RTT. In order to test our implementation, we utilized the ping command to generate one hundred ICMP packets from UE to the N6 interface of the UPF, which is responsible for the communication between the 5G network and the PDN, to obtain precise experimental results. As seen in Figure 3.3, the integration of WiFi DU supports the usage of WiFi DU for transmitting UE packets resulting in a very short RTT time for the packets compared to that of the 5G DU, regardless of the USRP and stream scheme used. The mean values of RTT for the WiFi DU, 5G-DU b210, 5G-DU n310 1x1, and 5G-DU n310 2x2 are 2.509, 11.752, 12.01, 12.04 ms, respectively. The results denote that the WiFi DU outperforms the tested 5G DUs for the RTT metric.

To continue evaluating the proposed system, we used another key metric of the 5G, the application DL throughput. The DL throughput evaluation is done by generating UDP traffic from the N6 interface to the UE via iperf. Figure 3.4 illustrates that the DL throughput

via the WiFi DU outperforms the 5G DUs since we managed to succeed at over 100 Mbps data transmission speed, where 5G-DU b210 managed to succeed at a speed of around 50 Mbps, and 5G-DU n310 1x1, and 5G-DU n310 2x2 around 60 Mbps. This performance in DL throughput highlights the significant contribution to the increased network capacity of this hybrid 5G system.

As far as the uplink transmission is concerned, the achieved transmission speeds differ to a large extent. 5G DU supports an average transmission speed of 3.95 Mbps, while with WiFi DU, 155 Mbps, achieving 39.24 times faster connectivity.

One of the challenges of integrating WiFi technology into the 5G architecture was the design of a slicing scheme as well as the co-existence of multiple clients. WiFi tends to distribute equally the available bandwidth among all active clients. This behavior is illustrated in Figure 3.5, where we conducted an experiment in which the same amount of UDP traffic was generated from the N6 interface to two UEs via iperf. In the initial phase of the experiment, the first 10 seconds, we exclusively generate traffic to station 1, allowing it to monopolize the channel and to succeed DL traffic speed of over 100 Mbps. In the second phase of the experiment, after 10 seconds, we also began to generate traffic to station 2, which leads to equal distribution of the medium since both stations tend to succeed a throughput of 50Mbps. When the transmission to station 1 ends after 10 seconds, station 2 monopolizes the channel (100 Mbps).

To evaluate our proposed WiFi slicing mechanism, we replicated the previous experiment with a key modification of assigning UEs to different user-based slices as described in Section 3.3.2 and illustrated in Figure 3.2. Station 1 is assigned to the priority slice with a ToS value of 0xb8, which maps to User Priority 5 and therefore to the Video (VI) access category. Station 2 is assigned to the non-priority slice with a ToS value of 0x04, which maps to User Priority 0 and therefore to the Best Effort (BE) access category. Since VI has shorter inter-frame spacing and a smaller contention window than BE, Station 1 gains more frequent access to the channel. As shown in Figure 3.6, in the first phase of the experiment Station 1 monopolizes the channel, achieving DL throughput of 100 Mbps. The channel's resources are unequally divided when station 2 started to receive traffic since the clients are now assigned to different ACs with different opportunistic channel access characteristics, giving different channel access priorities. Instead, station 1 continued to operate at a higher throughput of 75 Mbps, while station 2 only managed 37 Mbps. This means that station 1 used about 70%

of the channel capacity, which is a significant increase from the 50% seen in the non-sliced experiment. Prioritization continued until the end of station 1's transmission, at which point station 2 began to monopolize the channel achieving over 100Mbps.

Based on our results, we can see that integrating WiFi DU into 5G network offers a significant reduction in the RTT time for UEs, as well as improvements over the transmission data speed resulting to increased network's capacity. WiFi vulnerabilities regarding the coexistence of multiple clients are solved by our slicing mechanism in which we can share the medium according to their demands.

3.5 Conclusions

In this chapter, we have implemented a multi-RAT 5G network architecture by integrating a WiFi DU into real-world 5G networks using the OAI RAN. Our proposed system architecture leverages Software-Defined Networking (SDN) and the ToS field of IP packets to implement priority-based (airtime) slicing among clients. We conducted experiments in RF-isolated real-world environments in order to demonstrate the viability of our proposed architecture. The results indicate that the WiFi integration offers the lowest packet latency and the highest data throughput among the tested 5G DU deployments, and thus can be used to augment network capacity and provide low-latency services over the non-3GPP DU.

However, a limitation of this chapter is that the traffic distribution across the multi-RAT DUs relies on static policies. In a real-world deployment, traffic demand fluctuates, requiring more sophisticated, dynamic traffic steering to better exploit the complementary characteristics of multiple RATs. This dynamic control can be realized through the O-RAN RICs, in order to enable real-time adaptation of the RAN to the UE's demands. These two directions — dynamic traffic steering and O-RAN integration — form the focus of the following chapter.

Chapter 4

Proactive, ML-Driven Traffic Steering in Multi-RAT O-RAN Systems

4.1 Introduction

Building upon the hybrid 5G-WiFi architecture presented in Chapter 3, this chapter introduces a methodology to efficiently utilize heterogeneous access resources in a multi-RAT environment through traffic steering. In contrast to Chapter 3, which integrates WiFi as a non-3GPP DU within a C-RAN architecture relying on static traffic distribution policies, this chapter extends the previous approach by integrating WiFi as a non-3GPP DU in an O-RAN architecture, and proposes traffic-steering policies on how traffic can be dynamically distributed across the available RATs under different network conditions. Addressing this limitation requires a mechanism capable of dynamically distributing traffic across the available RATs, a problem that the 3GPP has also recognized. The key difference between the approach in Chapter 3 and the 3GPP-standardized solution lies in the point of integration of WiFi technology into the 5G system. As already mentioned, 3GPP integrates WiFi at the 5GCN through the gateway functions of N3IWF and TNGF. 3GPP's solution for traffic steering, according to the 3GPP Release 16 [48] is achieved via the introduction of a new CN component, namely the Access Traffic Steering, Switching, and Splitting (ATSSS) function to enable multi-access traffic handling across 3GPP and non-3GPP links, primarily through the use of Multipath TCP (MPTCP) at the transport layer. On the contrary, this thesis designs and implements a tighter integration approach in which WiFi is embedded directly at the RAN level as a non-3GPP DU, managed by the CU through the F1 interface, and therefore

requires a RAN-oriented traffic steering mechanism.

Traffic steering can enable efficient network operations, and its importance grows significantly in the 6G era, where the UC service category envisions seamless connectivity offered by a heterogeneous mix of RATs such as 5G NR, WiFi, and satellite, where intelligent and proactive steering across these interfaces will be a cornerstone of AI-native network management. The integration of AI into traffic steering decisions is necessary because, in a multi-RAT network environment, static traffic steering does not fully leverage RAT characteristics such as capacity, latency, and channel conditions, and is limited to simple steering policies such as round-robin, weighted scheduling, and others. Using AI/ML, traffic steering decisions can incorporate demand and network condition forecasting, enabling proactive distribution of traffic across available RATs based on predicted network conditions, in line with the principles of 6G's IAC service category. This raises the **RQ2**: **How can an AI-driven control mechanism within the O-RAN architecture enable proactive and intelligent traffic steering in a heterogeneous multi-RAT system?**

From a design point of view, the RAN part of cellular networks requires two distinct components to support dynamic traffic steering. The first component is a controller responsible for monitoring radio conditions and, according to specific policies, making steering decisions. The second required component is a scheduler that will be placed within the 5G RAN components to implement the controller's steering decisions. The O-RAN architecture, introduced in Section 2.2.2, provides the near-RT RIC, which acts as the controller, collecting KPM telemetry in real time from the subscribed E2 nodes and executing xApps that implement the steering logic and make steering decisions. Since we are based on WiFi integration in the 5G-RAN, as described in Chapter 3, the missing scheduler component can be implemented at the CU level of the disaggregated RAN, and packets will be forwarded to the selected RAT-DU based on decisions made by the xApp.

Specifically, this chapter designs and implements a multi-RAT O-RAN system where WiFi is integrated as a non-3GPP DU at the MAC layer of the 5G RAN stack, managed by the CU through the F1 interface, and introduces a proactive xApp deployed on the Near-RT RIC that utilizes a pre-trained Long Short-Term Memory (LSTM) model to forecast UE throughput demands and performs dynamically packet-level steering decisions to a custom CU-level scheduler, which distributes traffic across the 5G NR and WiFi DUs based on the predicted load, enabling demand-aware multi-RAT traffic management within the O-RAN

AI/ML workflow established in Section 2.3.

Putting everything together, this Chapter attempts to:

- Design and implement a multi-RAT O-RAN system with WiFi DU integration.
- Design and implement an xApp for dynamic traffic steering in a multi-RAT 5G system, utilizing UE traffic load predictions.
- Experimentally evaluate the proposed mechanism in a real-world testbed and traffic environment

4.2 Related Work

The literature's solutions for integrating non-3GPP access technologies into 5G systems, along with the different standardization efforts around LWA, LWIP, N3IWF, and TNGF, are presented in Chapter 3. This Chapter's related work aims to present papers that specifically target traffic steering mechanisms deployed in 5G networks across homogeneous or heterogeneous RATs.

A real-world implementation and evaluation of the ATSSS is shown in [49], where the three defined mechanisms in [48], namely Switching (failover between links), Splitting (load distribution per flow), and Steering (policy-based path selection) using static routing rules, are deployed and evaluated. Several works explore the ATSSS mechanism for latency-aware enhancements over heterogeneous 5G and WiFi links [50] or DRL-based enhancements for predictive flow steering between 5G and WiFi access [51].

Multiple works aim to address traffic steering in 5G networks, targeting both homogeneous and heterogeneous RAT scenarios. Focusing on homogeneous RAT traffic steering, the authors in [52] propose a mechanism in which user flows are routed across 5G RUs using LSTM-based demand prediction and dynamic resource allocation via the O-RAN near-RT and non-RT RICs, and evaluate this approach in a simulated environment. A similar approach has been followed in the paper [53], where an xApp is responsible for performing traffic steering and load balancing via handovers, with decisions triggered by K-means clustering of UE throughput and QoS metrics to identify underperforming UEs, combined with LSTM-based cell throughput prediction. A user-specific traffic steering approach has been presented in the paper [54], where an xApp based on Q-learning and a CNN utilizes handover to optimally as-

sign each UE to a serving base station in a large-scale simulated 5G deployment, performing traffic steering at the homogeneous RAN level.

Regarding multi-RAT traffic steering, the work [55] proposes a hierarchical reinforcement learning mechanism that leverages the O-RAN architecture to handle different traffic types (e.g., video, voice, gaming) with distinct QoS needs by adjusting flow decisions in real time between LTE and 5G. In [56], the authors propose an ML-based mechanism that integrates LTE and WiFi DUs in a disaggregated 5G RAN, using a BiLSTM model to predict per-UE CQI and dynamically route traffic in real time via a FlexRAN-based xApp, based on predicted link quality toward the different RATs, with implementation and validation conducted in a real-world testbed environment.

This chapter differs from related work in three key aspects. First, unlike ATSSS-based approaches that operate solely at the CN level, the proposed mechanism performs traffic steering at the RAN level via a custom CU-level scheduler. Second, this chapter is not based on homogeneous RATs; instead, it proposes a concurrent multihoming UE connected to both 5G and WiFi DUs, with dynamic packet-level splitting orchestrated by an xApp. Third, unlike existing works that rely on simulation-based evaluations, this chapter validates the proposed mechanism on a real-world O-RAN testbed using real traffic traces.

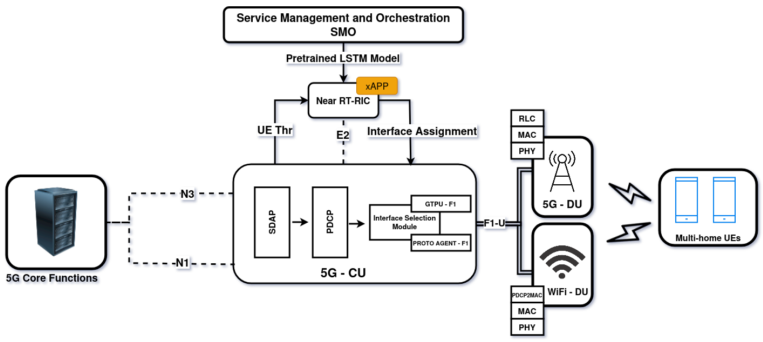


Figure 4.1: Proactive scaling O-RAN system architecture

4.3 System Architecture and Environment Setup

In this Section, we present the proposed system architecture that implements a hybrid 5G O-RAN that integrates WiFi as a non-3GPP DU at the MAC layer of the 5G protocol stack. In our base architecture, we consider the 5G network stack as a reference, following the 3GPP Option 2 split as presented in Section 2.2.2, in which the CU handles the PDCP layer and above, while the DU is responsible for the RLC layer and the layers below. The CU, 5G DU,

and WiFi DU are the three main components of this disaggregated Multi-RAT 5G O-RAN architecture. In the following subsections, we present the design and implementation of the CU and WiFi DU entities, and explain how an xApp can enable hierarchical traffic steering among the available DUs.

4.3.1 System Architecture

The system design is based on the O-RAN architecture and aims to provide flexibility, openness, and intelligent traffic management across a 5G multi-RAT network. The proposed 5G system architecture is based on a basic 5GCN, which includes the NRF, UDR, UDM, AUSF, AMF, SMF, and a software-based UPF component. It follows a standalone deployment with virtual functions, according to the Service-Based Architecture, as described in Section 2.2

Focusing on the RAN, the CU hosts the SDAP and PDCP layers and the DU includes the RLC and the layers below. Our 5G Multi-RAT RAN consists of a 5G DU with the typical NR MAC and PHY layers and a WiFi DU where these layers are replaced by the corresponding IEEE 802.11 MAC and PHY. This design allows the CU to manage both 5G and WiFi traffic in a unified way, without the need for additional interworking functions such as N3IWF. Scheduling can be handled at the PDCP layer, with a finer granularity, even in a per-packet basis. 5G and WiFi use different MAC policies; the 5G MAC primarily relies on centrally scheduled access, while WiFi MAC is contention-based, relying on CSMA/CA.

In this chapter, the WiFi DU is integrated into the O-RAN architecture as a MAC-level bridge between the CU and the WiFi radio, without a slicing mechanism. Although this integration was described in detail in Chapter 3, we briefly recall its main steps here to keep the present chapter self-contained. To support integration, in the DL case, PDCP packets are received from the CU and stripped of their PDCP headers before being injected into the wireless network. To support seamless integration of the WiFi DU into the 5G architecture, a mapping layer is introduced to map 5G RNTIs to WiFi MAC addresses, ensuring that the CU can correctly identify UEs across both RATs. The RNTI is extracted, and the WiFi-DU maps it to the corresponding MAC address using the mapping table. The DU subsequently adds the 802.11 MAC header and sends the frame through the WiFi NIC towards the UE. For the uplink case, the inverse process is performed. This ensures that the packets received by the CU are always bearing the PDCP header, while the CU always sends PDCP-encapsulated

packets to the DU, regardless of the RAN technology.

The midhaul between the CU and the 5G DU uses standard F1AP over GTP-U. For the WiFi DU, a custom F1-over-IP protocol operates over TCP to ensure ordered and reliable delivery of control and user data. The F1oIP handles DU registration and capability exchange so the CU can manage both DUs in a unified way.

The CU is responsible for managing both the PDCP sequence number aggregation and the packet scheduling across the 5G and WiFi DUs. Each DU maintains its own PDCP sequence numbering, and the CU centrally manages the network sequence numbers on a per-UE basis to ensure a consistent PDCP context and ensure packets are delivered in order to the 5GCN, even when uplink traffic is split between the DUs. Static rules or round-robin schedulers, as described in [45], lack the flexibility needed to adapt to rapid changes in channel quality, congestion, or evolving QoS requirements. In this chapter, we extend the implementation from the previous Chapter 3 by introducing a scheduler between PDCP and RLC at the CU, which dynamically determines how to split traffic between the 5G DU and the WiFi DU based on xApp decisions. The scheduler computes the packet ratio for each DU at each scheduling interval and ensures that traffic is split according to predicted demand and target WiFi ratios, allowing the system to adjust to current load conditions.

The system is O-RAN-based and supports intelligent traffic steering via a custom xApp running on the near-RT RIC. The xApp implements hierarchical steering logic, in which the 5G DU serves as the primary interface, and the WiFi DU is activated when ML-based demand forecasting indicates that 5G capacity will be exceeded. When the total traffic cannot be served by a single DU, either the 5G or the WiFi DU, the traffic is split between them. The xApp communicates over E2 to enforce the decisions at the CU. The SMO provides the pre-trained ML model used for demand forecasting, which drives traffic steering decisions.

4.3.2 xApp Functionality

The proposed xApp implements a non-prioritizing multi-RAT scheduling with a demand-aware WiFi allocation scheme. The goal is to distribute traffic across 5G and WiFi DUs without UE prioritization nor slicing support. The xApp works in three phases. First, it connects to the near-RT RIC, retrieves the pretrained ML model from the SMO, and subscribes to the E2 agent to receive the DL throughput KPIs (DRB.UEThpDl) from the CU, the only E2 node in the system.

Algorithm 1 xApp-Fair Share Multi-RAT

Require: constants TOTAL_5G_CAPACITY, TOTAL_WIFI_CAPACITY
Require: pretrained LSTM model
Phase 1 – Monitoring: Receive current $thrDl[u]$ from E2 interface
Phase 2 – Prediction: prediction[u] ← LSTM_forecasting($thrDl[u]$)
FAIR_5G_SHARE ← TOTAL_5G_CAPACITY / #UEs
Phase 3 – Routing Decision:
for $i = 0$ to $prediction_step$ step s **do**
    wifiCandidates ← {}
    for each UE u **do**
        if prediction[u][i] > FAIR_5G_SHARE **then**
            add u to wifiCandidates
        end if
    end for
    totalWifiDemand ← $\sum_{u \in wifiCandidates}$ prediction[u]
    if totalWifiDemand ≤ TOTAL_WIFI_CAPACITY **then**
        for each u in wifiCandidates **do**
            wifiShare[u] ← prediction[u]
        end for
    else
        share ← TOTAL_WIFI_CAPACITY / $|wifiCandidates|$
        for each u in wifiCandidates **do**
            wifiShare[u] ← share
        end for
    end if
    for each UE u **do**
        if prediction[$u + i$] ≤ FAIR_5G_SHARE **then**
            route[u] ← 5G_DU_only; wifiRatio[u] ← 0
        else if $u \in$ wifiCandidates and prediction[$u + 1$] ≤ wifiShare[u] **then**
            route[u] ← WiFi_DU_only; wifiRatio[u] ← 1
        else
            route[u] ← Split_5G+WiFi
            wifiRatio[u] ← (prediction[u] − FAIR_5G_SHARE) / prediction[u]
        end if
    end for
    send routing decisions, wifiRatio and forecasted throughput to CU via E2
end for

In Phase 1, the xApp monitors the reported DL throughput of the UEs. In Phase 2, at each T- interval, it uses the LSTM model to predict the UEs' throughput for the next two steps, based on the last 10 values. In Phase 3, it decides on the steering. The 5G DU is always treated as the main interface, and WiFi DU is activated only when needed.

To make effective traffic steering decisions, it is essential to model the behavior of both the 5G and WiFi channels. In the case of 5G, the system implements fair sharing of capacity via a basic scheduler that allocates available Physical Resource Blocks (PRBs) equally among

all connected UEs during each Transmission Time Interval (TTI). We model this by dividing the total DU capacity by the number of active UEs, resulting in an equal share of bandwidth per UE.

For WiFi, we adopt a CSMA/CA-based channel model. When the total demand is below the available capacity, the channel allocation may be uneven, as UEs use only the bandwidth they need. However, once the channel becomes saturated, a fair-sharing mechanism is applied. If the total demand exceeds what either of the available DUs can satisfy, the traffic is divided between them. After determining the percentage of traffic that each DU will transfer, the xApp sends the routing instructions and the WiFi packet ratios to the CU scheduler.

Algorithm 1 shows the entire workflow. This modeling approach was selected for its lightweight nature, enabling real-time execution within the xApp, unlike more complex models that require significantly greater computational effort and are not suitable for near-real-time operation.

4.3.3 The Necessity of ML-Based Demand Forecasting

Throughput prediction is an essential addition to our system, as it enables proactive traffic steering across the Multi-RAT DUs. Without forecasting, the scheduler would be limited to a reactive approach, responding only after congestion has occurred, resulting in a system that operates with suboptimal resource utilization and degrades the QoE offered to the UEs. Using an LSTM model, we can forecast future UE demand, so our xApp can proactively make decisions about the scheduler's forwarding rules before the multi-RAT DUs reach their capacity limits. We adopt the LSTM-based forecasting over statistical methods due to its ability to capture non-linear, time-varying UE traffic patterns common in mobile networks. The use of the LSTM improves scheduling efficiency by enabling traffic-steering decisions ahead of congestion. More details regarding the LSTM architecture used in our implementation can be found in Section 4.3.4.

4.3.4 Real World Implementation

Deployment Setup

Our deployment uses a cloud-native 5GCN implemented with OpenAirInterface (OAI), supporting the basic NFs: NRF, UDR, UDM, AUSF, AMF, SMF, and a software-based UPF.

All these 5GCN NFs are orchestrated using Docker Compose. Regarding the 5G RAN, we use a disaggregated architecture comprising a CU and two DUs: a 5G NR DU and a WiFi DU. The implementation of the 5G RAN uses bare-metal deployments. We use an OAI 5G DU without any code modifications and rely on the OAI RF simulator to emulate the radio link between each UE and the 5G DU, instead of establishing it over the air. This setup emulates realistic wireless channel conditions with an Additive White Gaussian Noise (AWGN) model, effectively replacing the physical RF boards. To deploy the 5G CU supporting the WiFi DU integration, we applied the required OAI code modifications as described in Section 4.3.1. The WiFi DU functionality was also designed and implemented through a Python-based agent that interfaces a local WiFi Access Point. This agent follows the approach designed in [46], [45], developed using the Google Protocol Buffers library to enable messaging exchange between the 5G CU and the WiFi DU. The WiFi DU uses the underlying AP implemented with the *hostapd* service, using the IEEE 802.11ac protocol at 5 GHz with a 40 MHz channel width and a single spatial stream, while the NIC used is the Qualcomm Atheros QCA988x, compatible with the *ath10k* driver. A Near-RT RIC was deployed using the OAI FlexRIC implementation and hosted our xApp. The experimental setup includes multi-homed UEs that connect to the WiFi DU using Qualcomm Atheros QCA988x cards and to the 5G DU via the RF simulator, enabling packet-level traffic steering across both interfaces.

Dataset

The dataset used to train our LSTM is described in Section 2.7. We retained only the DL bitrate feature and discarded the remaining features (e.g., location and ping metrics) as unnecessary. Data from three of the five available days was used to train the LSTM, while the remaining two days were used to generate traffic for two UEs. Preprocessing included min-max normalization in the range of 0–100 Mbps and spline interpolation to smooth the data and reduce noise.

LSTM

To forecast UEs' throughput demands, we used the popular time-series forecasting method, namely the LSTM neural network, which is widely used for cellular traffic prediction as seen in [57, 58]. The model includes two LSTM layers (224 and 64 neurons) and a dense output layer with two units. Its architecture and hyperparameters were selected using Keras Tuner,

and the model was trained offline on the SMO O-RAN component for 10 epochs with the Adam optimizer (learning rate 0.0034). The LSTM achieved an MAE of 1.384 on the test set, balancing accuracy and generalization. Single-step forecasting was too short-sighted for the O-RAN control loop, whereas a two-step horizon proved the most effective, outperforming other architectures with longer input windows or prediction horizons.

4.4 Experimental Environment & Evaluation

To evaluate the proposed multi-RAT 5G system under realistic conditions, as mentioned in Section 2.6.1, we conducted extensive experiments on the NITOS testbed, the Greek node of SLICES-RI. NITOS provides a controlled and flexible environment that enables real-world experimentation with cutting-edge wireless technologies, including 5G and WiFi. These experiments were designed to demonstrate the viability of the proposed O-RAN multi-RAT system and the traffic steering mechanism.

4.4.1 Experimental Conditions

From experiments conducted on the NITOS testbed, several key observations were made. The maximum DL throughput achieved by the 5G DU was approximately 60 Mbps, while the WiFi DU reached up to 120 Mbps. The uplink results were also notable, with the WiFi DU delivering significantly better performance than the 5G DU. Specifically, the 5G DU uplink was limited to 10 Mbps, whereas the WiFi DU, benefiting from its symmetric design, provided uplink speeds comparable to its DL performance. Another important finding was the latency difference between the two DUs: the WiFi DU, with its simpler protocol stack, achieved an average RTT of approximately 2.5 ms, while the 5G DU reached approximately 11.8 ms. We should note that we run the experiments in an isolated environment, with no other transmissions in the same area. These experiments also defined the constants required by the xApp algorithm, namely the 5G capacity, set at 60 Mbps, and the WiFi capacity, set at 120 Mbps, which were used as reference values for traffic steering decisions during the evaluation.

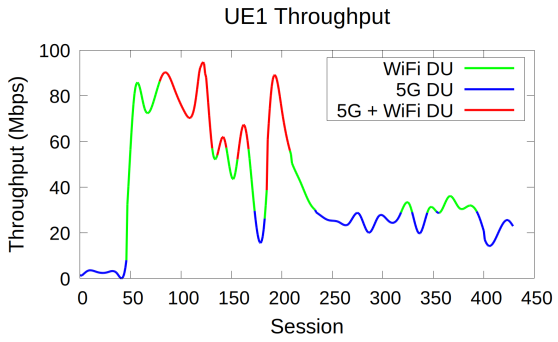


Figure 4.2: UE1 Throughput

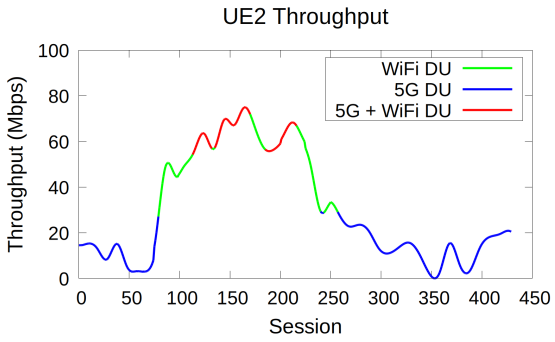


Figure 4.3: UE2 Throughput

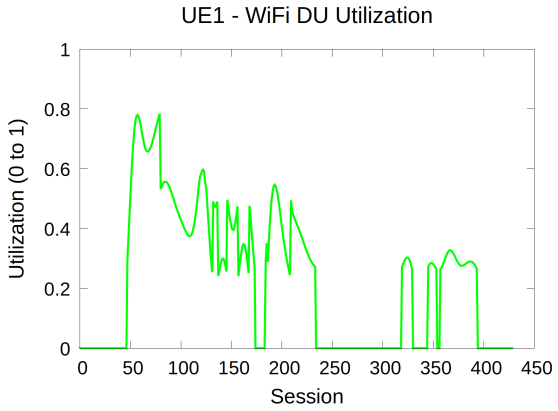


Figure 4.4: UE1 - WiFi DU Utilization

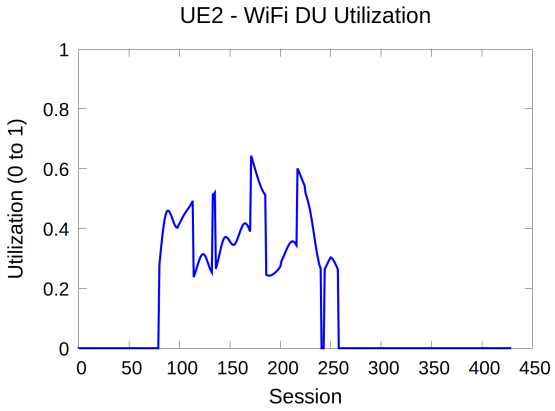


Figure 4.5: UE2 - WiFi DU Utilization

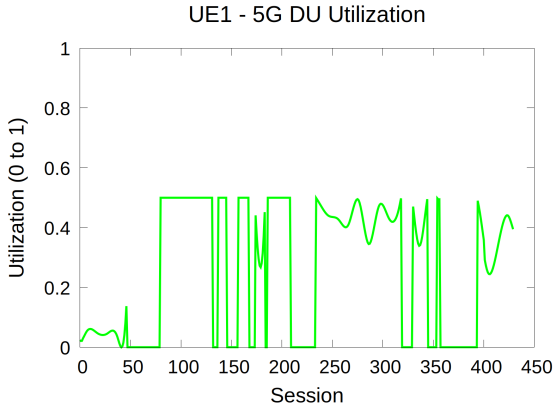


Figure 4.6: UE1 - 5G DU Utilization

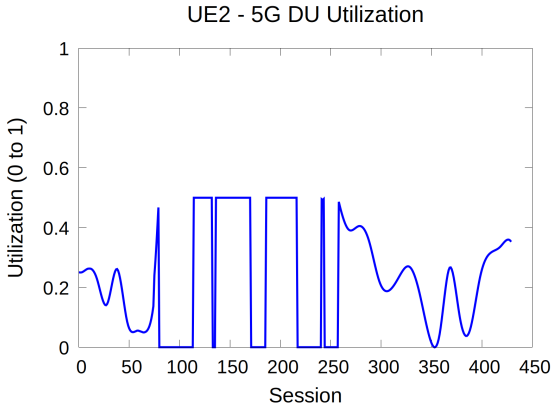


Figure 4.7: UE2 - 5G DU Utilization

4.4.2 Experiments & Evaluation

We evaluate the proposed system using real-world traffic traces, as described in Section 4.3.4. We utilize the iperf command to generate DL UDP traffic from the UPF's N6 interface to the UE. Figures 4.2 and 4.3 illustrates the per-session DL throughput for two UEs (UE1 and UE2), based on traffic measurements collected at the UE side.

The color-coded segments indicate the routing interface used during each session: WiFi

DU (green), 5G DU (blue), and simultaneous use of both (red). In the initial phase (sessions 0-≈ 50), the traffic generated for both UE1 and UE2 is relatively low. During this period, the traffic is routed entirely through the 5G DU, since 5G DU's capacity is sufficient to serve the predicted demand. The utilization of the 5G DU increases proportional to the UE's traffic while the WiFi DU remains inactive.

As traffic demand rises—particularly for UE1 around session ≈50, the xApp identifies that the 5G DU is no longer sufficient to support the demands of the UE1. Consequently, it begins routing UE1's traffic through the WiFi DU. This behavior is clearly visible in Figure 4.2 (transition to green and red segments for UE1) and in Figure 4.4, where a sharp increase is observed.

A notable transition occurs around session ≈75, where UE2's demand increases significantly, and the 5G DU becomes insufficient to serve it. In response, the xApp begins routing UE2's traffic over the WiFi DU. This leads to a rise in WiFi DU utilization as seen in Figures 4.4 and 4.5, creating a contention environment in the WiFi DU where the UE1 traffic must be distributed across both interfaces in order to avoid WiFi contention. Since the non-prioritizing xApp applies no user-specific preference, both UEs are treated equally, meaning both UEs are able to utilize concurrently almost the same percentage of the DUs, and UE1 adapts by dynamically offloading part of its demand between the 5G and WiFi DUs.

From this point onward, UE1 maintains dual-interface operation for an extended period, which aligns with its higher predicted throughput compared to UE2. This behavior is consistently reflected in the WiFi and 5G DU utilization plots, where UE1 demonstrates more persistent interface usage, whereas UE2 transitions back to single-interface operation earlier due to its lower demand. After the ≈200th session, the overall traffic demand decreases for both users. This is evident from the declining throughput values and the reduced utilization on both interfaces.

The system is shown to adaptively and dynamically steer traffic based on predicted demand, leveraging the full potential of both the 5G DU and WiFi DU interfaces. The fairness of the xApp is achieved since both UE1 and UE2 receive proportional access to resources, and no prioritization logic distorts the allocation. Even in high contention scenarios, traffic is balanced and interfaces are shared dynamically without degradation in throughput performance.

4.5 Conclusion

In this chapter, we presented the design, implementation, and experimental evaluation of a multi-RAT 5G O-RAN system that integrates WiFi as a non-3GPP DU. A custom scheduler at the CU was developed to enable packet-level traffic steering across multi-RAT DUs, while a non-prioritizing xApp was designed to dynamically make routing decisions from LSTM throughput predictions. The proposed system was experimentally validated under real settings using a real-world dataset. The results highlight the system's ability to dynamically adapt to traffic fluctuations and balance DU resources among UEs, ensuring fair treatment under varying load conditions. By using the near RT-RIC, we managed the resources of the 5G and WiFi DUs at the RAN level, demonstrating that traffic steering is not only feasible but also highly effective. The experimental results highlight the potential of O-RAN-based approaches to enable intelligent and demand-aware network control mechanisms in next-generation wireless systems.

The experiments also reveal another important consequence of the multi-RAT usage in the 5G networks, that as the multi-RAT utilization increases, the UP components of the RAN, for instance the CU-UP can become a critical bottleneck since the aggregated traffic volume processed by the CU-UP grows significantly. This raises a critical research question: can the CU-UP dynamically scale to meet this growing demand, and can a proactive mechanism be designed to ensure it does so before congestion occurs? This is precisely the question addressed in the following chapter.

Chapter 5

On the feasibility of Dynamic Scaling of the 5G O-RAN User Plane

5.1 Introduction

Following the design and control mechanisms introduced in Chapters 3 and 4, which enhance the capacity and utilization of the 5G RAN by integrating multi-RAT and dynamic traffic steering, this chapter addresses a key limitation of these approaches. While these mechanisms improve the effective use of radio resources by offloading traffic across heterogeneous DUs and dynamically distributing the traffic demand, they also introduce increased demands on the 5G RAN UP. As described in Section 2.2.2, the CU-UP is responsible for PDCP ciphering, SDAP marking, packet routing, header compression, and the aggregation of traffic from multiple DUs. The integration of multi-RAT DUs into 5G RAN increases the aggregated UP traffic volumes processed by the CU-UP, making the CU-UP a critical bottleneck that requires dynamic adaptation to sustain performance and reliability. This bottleneck defines the core research question of this chapter: **RQ3**: **How can cloud-native RAN UP functions be autonomously scaled within an O-RAN deployment to sustain service performance under dynamic and unpredictable traffic conditions?**

This chapter designs and implements a methodology to investigate the feasibility of RAN UP scaling in a fully cloud-native telecom system. In contrast to prior approaches introduced in [4], where a reactive scaling mechanism was developed in a V-RAN environment using platform-level KPIs (e.g., CPU utilization), this chapter extends that approach to a proactive, fully O-RAN-compliant framework.

To address the UP bottleneck in the 5G RAN, a mechanism that dynamically adapts CU-UP resources and resulting capacity in response to fluctuating traffic demands is required. The cloud-native and disaggregated characteristics of the V/O-RAN architecture make this feasible: the CU-UP is deployed as a VNF and can be horizontally scaled by deploying additional CU-UP instances across the available infrastructure nodes, distributing the aggregated UP load across the multiple active instances.

In contrast to the approach used in generic cloud-native scaling of stateless web services, where the standard k8s suite of autoscalers can scale by simply replicating a pod, scaling a stateful telecom VNF is much more complex. For instance, CU-UP maintains active PDCP sessions per UE, holds a unique CU-UP ID, and must be registered with the CU-CP via the E1 interface. This requires configuration updates and UE reassignment upon every scaling event, so a new configuration update must be provided to each newly deployed CU-UP instance.

The timing of scaling decisions is critical. Reactive scaling incurs a period of overload between when congestion is detected and when the new CU-UP instance becomes operational, during which packet loss occurs and QoS degrades. Reactive scaling is also more susceptible to ping-pong effects, where short-lived traffic fluctuations trigger an unnecessary scale-in followed immediately by a scale-out, or vice-versa, introducing control and scaling overhead. A proactive scaling approach addresses both limitations by initiating scaling operations before congestion events occur, rather than reacting to them.

The O-RAN architecture provides a platform for proactive CU-UP scaling, as the Near-RT RIC continuously collects standardized KPIs from the CU-UP E2 nodes via the E2 interface, thereby providing the required monitoring layer to drive scaling decisions. Additionally, by deploying xApps that can apply ML operations to forecast future network traffic demands, following the O-RAN AI/ML workflow described in Section 2.3, the system can take proactive CU-UP scaling decisions and, via the O-RAN framework, trigger the corresponding scaling operations, without relying on any platform-level instrumentation outside the O-RAN control loop.

In this chapter, we deploy a fully disaggregated 5G O-RAN network that supports the separation of the CP and UP with the existence of a near-RT RIC, alongside a service-based 5G CN, in a fully cloud-based deployment utilizing the state-of-the-art Kubernetes (k8s) service orchestrator. We developed an LSTM-based ML model using a real-world dataset to forecast future UE load demands and designed and implemented a proactive CU-UP scaling mech-

anism based on these predictions. Unlike approaches limited to simulation environments, we evaluate our proposed mechanism under real-world traffic conditions and environment, demonstrating its practical applicability using a testbed infrastructure.

The main contributions of this chapter are the following:

- Design and implement an xApp for dynamic horizontal CU-UP scaling in a 5G O-RAN system, leveraging forecasting of UE traffic load.
- Validate the proposed xApp and scaling mechanism using real-world traffic in a testbed environment.
- Compare reactive and proactive CU-UP scaling mechanisms, highlighting the advantages of proactive scaling.
- Provide a proof-of-concept on the horizontal scaling mechanism for the RAN.

5.2 Related Work

To introduce a proactive scaling mechanism to a 5G O-RAN system, it is essential to first examine the methods and strategies for scaling, especially in cloud-based network systems. In this context, VNF scaling has played a major role, since it enables dynamic resource allocation by adjusting the capacity of virtual network functions to meet fluctuating network demands. Authors in [59] and [60] perform proactive VNF scaling based on the forecast client demands using online learning and optimization techniques. Their methods minimize costs and improve overall service quality in dynamic simulated NFV environments. In [61], the authors describe a VNF scaling process based on forecast UE demands comparing three ML models ARIMA, XGBoost, and LSTM, using a real-world dataset to emulate network load from 4G base stations, without applying this mechanism to a specific 5G NF.

Notable work has been done on AMF scaling in 5GCN. Authors in [62] introduce a CNN-LSTM model to predict registration requests for proactive scaling of the AMF component. They evaluate the proposed methods using real-world data in a simulated environment, enabling efficient and proactive resource management, with a focus on prediction accuracy and minimizing training duration. In [63], the authors implement a dynamic horizontal scaling mechanism for the AMF while considering factors such as the number of connected UEs, the number of active AMFs and resource allocation information. Using a real-world dataset,

the authors train a WaveNet Deep Neural Network (DNN) to classify system states based on demand and decide on scaling horizontally the AMF.

UPF scaling has also been discussed in the scientific community for its potential benefits, as it is the only component in the 5GCN user plane. In [64], the authors present a mechanism for dynamically managing 5GCN functions, focusing on UPF scaling based on real-time monitoring of node CPU usage. This work aims to prevent network congestion, enhance network performance, and provide valuable techniques for managing resource allocation. In [65], the authors propose an auto-scaling UPF system using real-time traffic demand monitoring, named Bit rate Aware Auto Scaling (BAAS), relying on simulations and using a reference architecture based on LTE rather than a native 5G framework. Authors in [66] investigate vertical, horizontal, and hybrid scaling strategies for VNFs of a 5G Network, such as the UPF and application VNFs hosted on MEC servers. Their findings highlight that vertical scaling optimizes resource usage, while horizontal scaling improves the system's scalability and reliability. To balance these benefits, a heuristic algorithm is developed, which implements a hybrid strategy that dynamically selects between vertical and horizontal scaling based on the requirements of each scenario. In [67], a real-time scaling mechanism for VPP-based UPFs is proposed. The work uses the VPP rate, which reflects the number of packets processed at once, as a key indicator. This mechanism relies on a system that monitors and analyzes traffic to adjust the number of UPFs as needed. Work [68] presents an AI-driven system that dynamically scales the UPF component using a Deep Reinforcement Learning (DRL) framework based on real-time traffic demands, ensuring QoS while minimizing CPU and power usage.

Scaling 5G-RAN VNFs has not been extensively explored. Specifically focusing on the RAN, an interesting study by Mudvari et al. [69] presents an ML-driven dynamic scaling approach for the user plane components of 5G. The authors utilize four ML models to forecast network load using a real-world dataset and dynamically implement a proactive scaling mechanism that adjusts either the resources of the monolithic gNB (vertical scaling) or the instances of the UPF (horizontal scaling). A notable study [70] focuses on RAN components and implements a methodology to illustrate the scaling requirements for the CU, DU, and Near-RT RIC. The authors use an ILP model driven by forecast demands generated by two ML models, LSTM and ARIMA. The ILP model determines the optimal number of instances needed for the scaled components. However, their work is limited to traffic prediction and theoretical scaling strategies, as they do not implement or evaluate the scaling mechanism in

practice.

Our work differs from the existing literature by expanding research on scaling 5G infrastructure, moving beyond scaling of the 5GCN to focus on scaling actual user-plane RAN components. Contrary to prior work, we go beyond the state of the art and explore the benefits of horizontal scaling in the RAN. Existing works that have explored RAN scaling focus on the vertical scaling process (i.e., adding more computational resources to the function), contrary to our approach (spawning multiple replicas of the RAN function and directing traffic to them). By leveraging a real-world testbed, our work demonstrates the practical applicability of a proactive scaling mechanism for the CU-UP component, driven by clients' forecast traffic demands, in an O-RAN environment.

5.3 Empirical Analysis of CU-UP CPU Utilization and Traffic Load

This section presents an experimental analysis of the relationship between the incoming traffic load and the CPU utilization at the CU-UP, highlighting the impact on the system performance.

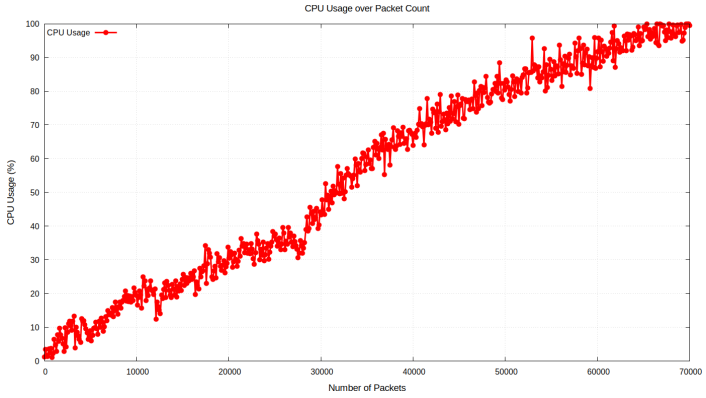


Figure 5.1: Correlation Between Incoming Traffic Load and CU-UP CPU Utilization

In the 5G RAN, the CU-UP behaves like a packet-processing pipeline. The CPU utilization of the CU-UP is highly correlated with the number of incoming packets due to the cost of per-packet processing, which includes reading packets from the NIC, header parsing, ciphering/deciphering, header compression, GTP-U encapsulation/decapsulation, and packet forwarding. The aforementioned relationship can be seen in Figure 5.1, where the number of

incoming packets increases while CPU usage rises accordingly. When the CPU utilization exceeds 90%, the CU-UP is not able to serve the incoming traffic, resulting in service degradation as observed in Figures 5.2 and 5.3. Specifically, in this experiment, a single CU-UP is deployed, and the same UDP traffic pattern shown in Figure 5.2 was generated from the DN to three connected UEs. As shown in Figures 5.2 and 5.3, when the traffic demands reach the highest peak (sessions 450 - 950), the CPU Usage of the CU-UP reaches 100%, and the received DL traffic at each UE is less than the requested rate due to packet loss caused by the high CPU utilization of the CU-UP. This reflects a QoS degradation due to the overloaded CU-UP.

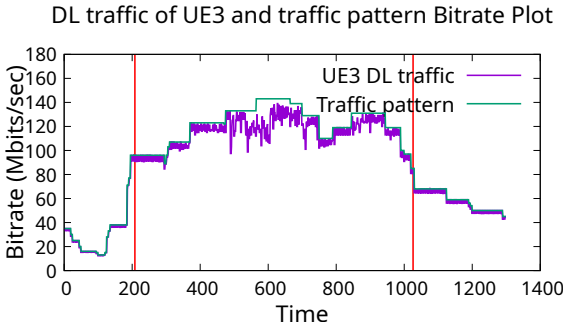


Figure 5.2: Downlink Traffic Profile Driving CU-UP Processing Demand

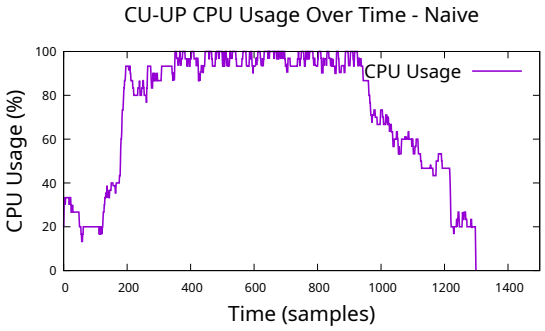


Figure 5.3: CU-UP CPU Utilization

Consequently, a system that monitors the incoming packet count of the CU-UP or its CPU Usage Utilization and performs horizontal scaling operations is a natural step toward mitigating QoS degradation. However, the challenge with this approach is that neither the incoming packet count nor CPU Usage Utilization is defined as an O-RAN compliant KPI. For this purpose, the system proposed in this chapter uses the incoming traffic throughput reported by the O-RAN KPI DRB DL UE, which serves as an indirect indicator of the number of incoming packets.

5.4 System Model and Architecture

This section provides an extensive analysis of all the components of a 5G network that supports proactive CU-UP scaling based on predicted UE traffic demands, detailing both the design of our proposed system architecture as depicted in Figure 5.4 and the experimental setup as shown in Figure 5.5.

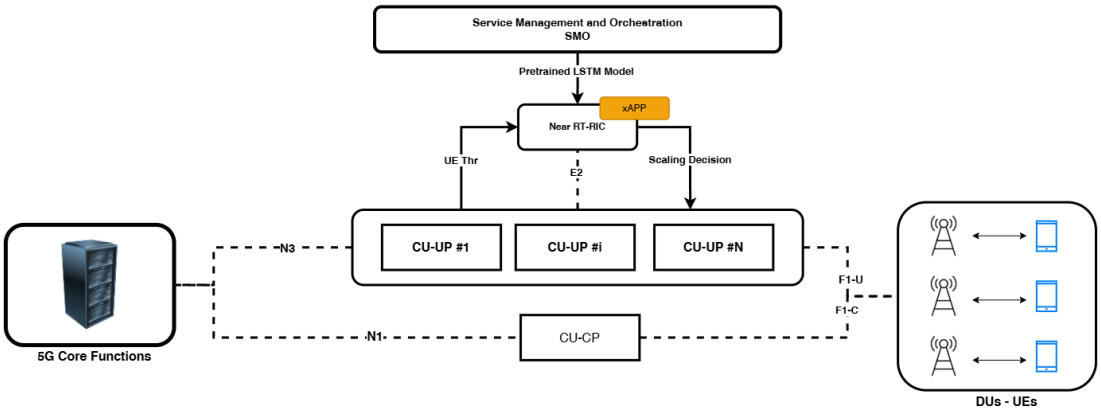


Figure 5.4: Proactive scaling O-RAN system architecture

5.4.1 5G Network Configuration and Setup

The proposed 5G system deploys both the 5GCN and the RAN in a cloud-based environment to enable virtualization, utilizing Kubernetes as the orchestration tool for these components. For our system architecture and experiment definition, we used the OAI software suite as mentioned in Section 2.5.1, which offers a software-based implementation of the 5GCN, RAN and RIC. For the 5GCN deployment, we deploy a basic 5G CN, as described in Section 2.2. In our experimental setup, we deploy two UPFs; as prior works have shown (e.g. [71]) the selection of the UPF can highly affect the overall throughput that the 5G network can achieve. As we are using the software-based UPF provided by OAI, which is reported to have lower throughput than hardware-based alternatives, we opt to deploy two UPFs to increase the overall capacity of the 5G network. By following this approach in our setup and using the disaggregated RAN deployment options, the CU-UP is now the only congestion point in the UEs' data paths under test.

The RAN follows the disaggregated Option 2 split introduced in Section 2.2.2, with the CU further separated into CU-CP and CU-UP via the E1 interface. As already described, the CU-UP is responsible for all the user-plane packet processing, becoming the target of the proposed scaling mechanism. Putting everything together, the components of our experimental 5G RAN network are the following: CU-CP, CU-UP that can be scaled by our algorithm, 2 DUs, and 2 UEs. Similar to the approach we follow for the 5GCN, we use multiple DUs to achieve higher overall network capacity and a single point of stress for the data-plane traffic: the CU-UP. The radio link between each UE and DU is not established over the air.

Instead, we utilized the RF simulator, which emulates the functionality of an actual Radio Frequency (RF) board. We opt to simulate real-world channel conditions using a model with Additive white Gaussian noise (AWGN). Given that the disaggregated 5G-RAN can also support further splitting of the DU into O-DU and O-RU, with the O-RU using real HW radios, our decision to emulate the radio part plays no significant role in the under-test system (the CU-UP) and only helps relieve potential bottlenecks and truly stress-test the system.

Since we base our system on the O-RAN architecture, all the RAN VNFs deployed in this 5G-RAN (CU-CP, CU-UPs and DUs) can operate as E2 nodes managed by the Near Real-Time Radio Intelligent Controller (Near-RT RIC). However, since the proposed mechanism relies exclusively on user-plane KPIs, only the CU-UP instances are enabled as E2 nodes in our deployment. The communication between the RAN VNFs and the Near-RT RIC takes place over the E2 interface, specified by the O-RAN Alliance through the E2 Application Protocol (E2AP) and its service models, such as the E2SM-KPM. The Near-RT RIC plays an important role in managing and optimizing the various functions of the RAN components by continuously collecting KPIs and the corresponding KPMs from the E2 nodes. In this chapter, we use the KPI *DRB.UEThpDl*. The metric is defined in the 3GPP TS 28.552 standard [72] and measures the DL bitrate per UE arriving at the CU-UP. For our system, and since the RAN part is based on OAI, we use FlexRIC as our controller [22]. The resulting system architecture used in our experiments is shown in Figure 5.5. Building on this foundation, we developed an xApp that continuously monitors the selected KPI from the CU-UP instance, and periodically (for our experiments, the period is set to 5 seconds) it forecasts the next 10 values using a window of the most recent 40 values. The predictions are based on a pre-trained LSTM model that is integrated into the xApp. This LSTM model was pretrained in the SMO, using the dataset described in Section 2.7 and resulting in the ML model in Section 5.4.3. Based on the predicted values, the xApp decides according to predefined thresholds on whether the CU-UP instances of the 5G-VRAN should be scaled-out, scaled-in, or remain unchanged. It is important to mention that the proactive mechanism is operating exclusively in the scaling-out scenario (spawning multiple replicas) as performing proactive scaling-in (reducing the number of replicas) could result in packet loss due to the system functioning with reduced resources. In Section 5.5.2, we also present experiments on a system that supports reactive scaling of the CU-UP instances. The key difference in the reactive implementation is that it does not forecast events; instead, the system continuously monitors client demands and

Table 5.1: Node Assignments and Specifications

Node	Components	CPU	Cores	RAM
1	NRF, UDR, UDM, AUSF, AMF, SMF, CU-CP, FlexRIC	Intel Core Processor (Skylake, IBRS)	4	8GB
2,3	UPF (Slice: oai & ims)	Intel Core Processor (Skylake, IBRS)	4	8GB
4,5	CU-UPs	Intel Core Processor (Skylake, IBRS)	4	8GB
6	DU, UE (Pair 1)	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz	6	16GB
7	DU, UE (Pair 2)	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz	6	16GB

triggers scaling only when real-time values exceed the predefined limits.

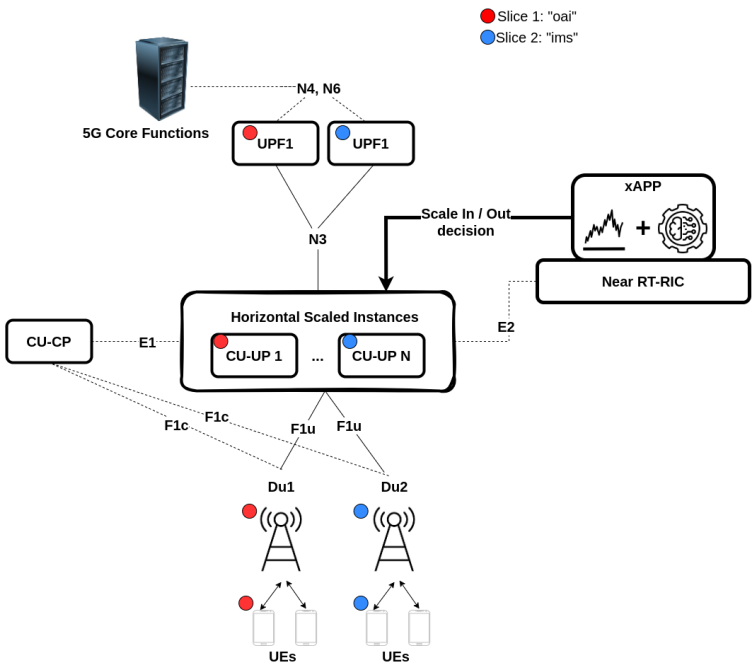


Figure 5.5: Experimental 5G CN and O-RAN architecture

5.4.2 Horizontal CU-UP Scaling Process

Horizontal scaling of network-based VNFs in real-world networks is not as straightforward as in cloud computing. The scaled CU-UP instances is not just about replicating the same functionality but instead creating new instances of the CU-UP components, even with different IDs, which need to be properly integrated and synchronized with the rest of the network components, thus requiring updates to their configurations. This is why traditional k8s tools for horizontal scaling, such as the Horizontal Pod Autoscaler, are unsuitable for scaling real-world 5G components, as they simply replicate the same service without making any configuration changes. As several of these services are stateful and require synchronization among the different components, special handling is needed to inform all involved components of the new replicas.

To scale the CU-UP in a seamless and efficient manner, the system must support the CU-UP Change functionality, allowing each UE to switch its assigned CU-UP without deregistering from the network. Although the CU-UP Change procedure is specified in 3GPP Release 15 [12], OAI has not yet released a version that implements this functionality. To address this limitation, we implement a scaling mechanism that prevents fault generation in the 5G network during scale-out events by first deploying the additional CU-UP instance and then restarting the corresponding UE instances assigned to it. For scale-in scenarios, we first uninstall the surplus CU-UP instances, then immediately restart the UEs connected to the uninstalled CU-UP instances to trigger reconnection to the 5G network and their assignment to the new ones. A limitation of this approach is the temporary interruption of UE connections caused by container restarts, and packet loss that may occur during transmission when the UE is disconnected. The CU-UP Change procedure, as outlined in [12], facilitates data forwarding from the source gNB-CU-UP to the target gNB-CU-UP, thereby preventing packet loss during the CU-UP switch. To eliminate packet loss in our implementation, we monitor the UE's connection state and ensure traffic is generated to/from the UE only when it is connected to the network. In production systems, these network readjustments and manual checks could be bypassed by adopting the CU-UP Change mechanism with data forwarding fully enabled. In our implementation, deploying new CU-UP instances requires negligible time, resulting in minimal impact on overall system performance.

Algorithm 2 Proactive CU-UP Scaling xApp

Require: Window Size W, Forecast Interval T, Prediction Horizon P
Require: Upper Threshold, Lower Threshold, Pretrained LSTM Model
Initialize: Sliding window $S[UE]$ with W historical DL throughput values per UE
Initialize: Real-time throughput $R[UE] \leftarrow 0$ for all UEs
Initialize: Load pretrained LSTM model
Initialize: Connect to RT-RIC and subscribe to the E2SM-KPM
while true **do**
  for each CU-UP (Parallel) **do**
    Fetch current DL throughput $R[UE]$ for each UE
    Update $S[UE]$ by appending $R[UE]$ and remove oldest entry
  end for
  if Every T seconds **then**
    Forecasting:
    Initialize $\text{Thr_Pr}_{\text{total}} \leftarrow 0$
    for all CU-UP (Parallel) **do**
      for all UE served by CU-UP (Parallel) **do**
        Predict the next P timesteps for each UE:
        $\text{Thr_Pr}[UE] \leftarrow \text{LSTM}(S[UE])$
        Compute $\text{Thr_Pr}_{\text{max}}[UE] \leftarrow \max(\text{Thr_Pr}[UE])$
        Update $\text{Thr_Pr}_{\text{total}} \leftarrow \text{Thr_Pr}_{\text{total}} + \text{Thr_Pr}_{\text{max}}[UE]$
      end for
      if $\text{Thr_Pr}_{\text{total}} \geq$ Upper Threshold **then**
        Append CU-UP to `Scale_OUT_list`
      end if
      if $\text{Thr_Pr}_{\text{total}} <$ Lower Threshold **then**
        Append CU-UP to `Scale_IN_list`
      end if
    end for
  end if
  Scaling Actions:
  Scale-Out:
  for all CU-UP in `Scale_OUT_list` (Parallel) **do**
    Deploy a new CU-UP
    Distribute UEs evenly across the existing CU-UP
    and the new CU-UP
  end for
  Scale-In:
  Create Pairs of CU-UP from the `Scale_IN_list`
  for all Pairs of CU-UPs in `Scale_IN_list` (Parallel) **do**
    Compute $\text{Thr_Pr}_{\text{total}}$ for the pair (CU_UP_i, CU_UP_j)
    if $\text{Thr_Pr}_{\text{total}} <$ Lower Threshold **then**
      for P timesteps **do**
        Initialize $R_{\text{total}} \leftarrow 0$
        for all UE served by the CU-UP pair (Parallel) **do**
          Fetch current throughput $R[UE]$
          Update sliding window $S[UE]$ with $R[UE]$ and
          remove the oldest entry
          Update $R_{\text{total}} \leftarrow R_{\text{total}} + R[UE]$
        end for

Algorithm 3 Proactive CU-UP Scaling xApp (Part 2)

```
48:                 if R_total < Lower Threshold then
49:                     Deallocate one CU-UP (CU_UP_i or CU_UP_j)
50:                     Assign the UEs from CU_UP_i and CU_UP_j to
51:                     the remaining CU-UP
52:                     break
53:                 end if
54:             end for
55:         end if
56:     end for
57: end while
```

5.4.3 Machine Learning Model - Demand Prediction

To forecast the future throughput values of the UEs, we replay a dataset as described in Section 2.7 over the network to generate traffic over the 5G network components. We collect the KPI values reported by the xApp and try to predict future demand across the network. We treat the specific problem as a time-series forecasting problem and choose to use the LSTM model, a popular model for forecasting the demand values of the UEs in 4G and 5G networks as illustrated in several works in relevant literature (e.g., [73], [74]). We utilized the same model as in Section 4.3.4, which consists of three layers: two LSTM layers with 96 and 128 neurons, respectively, followed by a dense output layer that predicts the next 10 time steps based on an input sequence of 40 time steps. The model architecture was generated using keras-tuner, which yielded optimal settings based on the training dataset and the number of future time points we wanted it to predict. The model was trained for 15 epochs using the Adam optimizer with a learning rate of 0.0015, balancing sufficient training value and preventing overfitting, achieving an accuracy of 5.6% MAPE. The primary purpose of this work is to study a 5G system that adapts the number of CU-UP instances based on the predicted throughput requirements of the UEs, rather than optimizing the prediction model itself. For this reason, we did not compare the performance of the resulting LSTM with other ML models. The trained LSTM model was saved and integrated as a pre-trained model into our xApp to predict the next 10 time steps in real time, using knowledge of the 40 previous values.

5.4.4 xApp Functionality

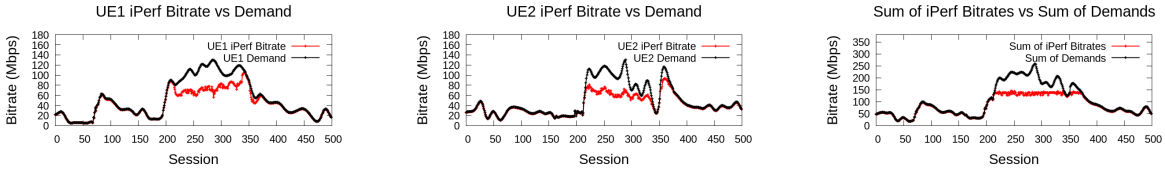
The proactive CU-UP scaling algorithm is designed to efficiently manage traffic loads in 5G O-RAN by forecasting future throughput demands and dynamically adjusting the number of CU-UP instances. This algorithm is implemented as part of an xApp deployed on the Near-RT RIC. Initially, the xApp retrieves a pre-trained machine learning model from the O-RAN SMO framework, connects to the Near-RT RIC, and subscribes to the E2SM-KPM offered by the E2 Agent. These entities exist in the FlexRIC implementation that we are using for this work. In this work, only the CU-UP nodes of the RAN are enabled as E2 nodes, since the target is to exclusively retrieve KPIs from the CU-UP components. The xApp retrieves the KPI *DRB.UEThpDl* from the E2 nodes every second and, at predefined intervals T (5 sessions in this case), it uses the pre-trained LSTM model to forecast the next predefined interval (10 sessions in this case) for each UE based on the previous W timesteps stored in a sliding window S. Then, it computes the total predicted throughput demand (Th_Pr_total) as the sum of the maximum predicted throughput values across all UEs per CU-UP. For each CU-UP that the computed total exceeds or equals a predefined upper threshold, the system scales-out by deploying a new CU-UP instance and redistributing the UEs evenly across the existing and newly deployed CU-UP. Conversely, if the total forecast throughput demand for the CU-UPs falls below a predefined lower threshold, the xApp initiates a scale-in procedure. During this process, it waits for the first value below the bottom threshold, then destroys the excess CU-UPs and redistributes the UEs among the remaining active CU-UPs. The functionality of the xApp is summarized in Algorithm 2.

5.4.5 Experimentation environment

To deploy our system in a real-world experimental environment, we utilized the NITOS testbed, the Greek Node of the SLICES-RI, hosted at the University of Thessaly, Greece as mentioned in Section 2.6.1.

During the design phase, we determined the number of nodes required for component deployment and assigned them to ensure an optimal configuration. The components were allocated as follows: a single node was designated for all 5G basic control plane entities, including NRF, UDR, UDM, AUSF, AMF, SMF, and CU-CP. These components, while latency-sensitive, have minimal computational demands. Two additional nodes were reserved for potential CU-UP deployments. Furthermore, two nodes with latency similar to that of the

CU-UP nodes (approximately 0.7 ms) were selected to deploy the two UPFs. Lastly, two additional nodes were allocated for deploying the two DU-UE pairs, which require significant computational resources and share a latency of approximately 0.5 ms with the CU-UP nodes. Table 5.1 provides a detailed overview of these assignments.

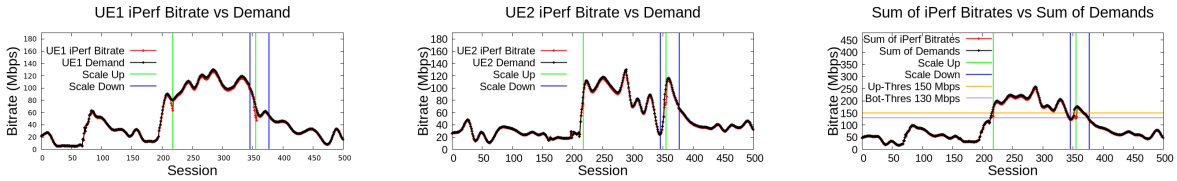


(a) UE1 iPerf Bitrate vs Demand

(b) UE2 iPerf Bitrate vs Demand

(c) Sum of iPerf Bitrates vs Demands

Figure 5.6: Naive System: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate.

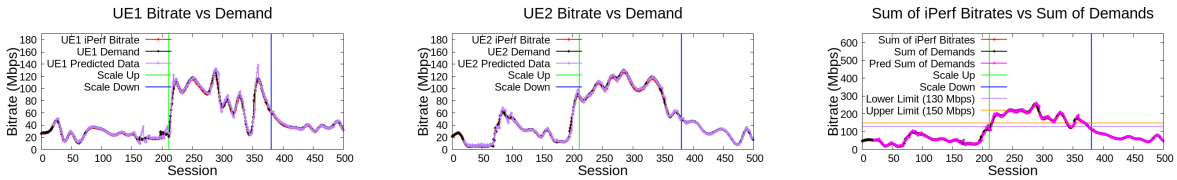


(a) UE1 iPerf Bitrate vs Demand

(b) UE2 iPerf Bitrate vs Demand

(c) Sum of iPerf Bitrates vs Demands

Figure 5.7: Reactive CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate.



(a) UE1 iPerf Bitrate vs Demand

(b) UE2 iPerf Bitrate vs Demand

(c) Sum of iPerf Bitrates vs Demands

Figure 5.8: Proactive CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate.

5.5 Evaluation

In this section, we present our experimental findings. Initially, we present the dataset we use to generate the corresponding network traffic, and subsequently the results and benefits our solution provides.

5.5.1 Dataset

The dataset used in this chapter is described in Section 2.7. Following the same approach as in Chapter 4, we retain only the DL bitrate feature, as it is the only metric relevant to the throughput forecasting task. Three of the five available day snapshots were used as the training dataset for the LSTM model, while the remaining two days were replayed over the network to simulate the traffic demands of the two UEs under test. During the preprocessing stage, min-max scaling was applied within the operating range of our system (0-150 Mbps), followed by spline interpolation to smooth the data and reduce noise, making it suitable for training the model. It is worth noting that the normalization upper bound differs from that in Chapter 4, where 100 Mbps was used, reflecting the 150 Mbps capacity limit imposed on the CU-UP instances in this chapter's experimental setup.

5.5.2 Experiments

To evaluate the performance of our system, we conduct three types of experiments. Initially, we study a naive system that does not support the CU-UP scaling mechanism. A second experiment is conducted to measure the performance of a system that supports the CU-UP scaling mechanism, based on real-time aggregation of UEs' throughput requirements via the xApp in FlexRIC. We call this method the reactive scaling method. At the same time, we evaluate the performance of a system that supports CU-UP scaling in and out, based on forecasted future requirements for scale-out and scale-in. In the latter case, we forecast when the network shall scale-in, and when that timepoint is reached, we scale the network. This allows us to continuously serve clients up to the point that they stop demanding traffic, and then scale-in the network, based on the real-time metrics that we receive. We call this method the proactive scaling method. After conducting various experiments, we found that the version of the OAI-CU-UP we use can meet clients' transmission requirements up to 300 Mbps. The maximum traffic that a CU-UP can handle depends on the degree of CPU utilization based on

incoming traffic, since other resources (such as RAM usage) remained underutilized during the experiments.

For the needs of this work, we limit the upper traffic boundary per CU-UP to 150 Mbps by assigning fewer resources to the scheduled pod. We argue that the usage of an upper limit for the maximum transmitted traffic of a CU-UP in our experiments does not violate the applicability of their results, since the performance characteristics of a CU-UP (v-RAN component) regarding maximum transmittable traffic are related to the assigned CPU resources [75]. Additionally, the upper and lower thresholds used in Algorithm 2 were set to 150 Mbps and 130 Mbps, respectively; these values can be adjusted based on the system configuration.

As mentioned in Section 2.7, two days of the dataset are used to simulate client requirements. In the presented experiments, the client transmission requirements are simulated by injecting UDP traffic from the UPF to the UEs using the *iperf* application. Each value of the testing dataset corresponds to a UDP session; for this reason, in Figures 5.7 and 5.8 no drop in the transmissions resulting from the scaling process, as described in Section 5.4, is visible.

Naive system

In Figure 5.6, we show the performance of the naive 5G network to satisfy the requirements of the connected UEs under different load conditions. As shown in Fig. 5.6a and 5.6b, the naive system can meet the needs of its connected UEs under normal conditions. However, the interesting part of the experiment is located between sessions 214 and 380 where the requirements of both UEs exceed 150 Mbps, which is the allowable traffic limit. Thus, the naive system is unable to satisfy the requirements of the UEs due to limitations in the CU-UP component's available capacity. This snapshot simulates peak times, where the increased demand from multiple users causes congestion. This case highlights the importance of dynamically scaling network resources through mechanisms such as adaptive bandwidth allocation that implies the activation of additional CU-UP instances, which can respond in real-time to changes in user demands.

Reactive CU-UP scaling system

In Figure 5.7, the graphs of the CU-UP Horizontal Scaling System based on real-time throughput monitoring under the same demand pattern are presented. As can be seen from Fig. 5.7a and 5.7b, this system is able to satisfy the demands of both of the connected UEs in

the case of normal conditions. At the same time, by scaling the CU-UP instances it can satisfy the total demands of the clients even when they exceed the configured limits. This scaling mechanism is responsible for preventing the severe performance degradation observed in the naive system. Vertical green and blue lines mark the moments when scaling-out or in of the CU-UP instances was performed, respectively. Although this system successfully handles all the demands of the clients even under high peak loads, we observe that relying to the real-time monitoring, introduces packet loss for a short period of time until the scale-out process is completed. This is happening due to the readjustment of the number of CU-UP instances, as it happens after the detection of the overlimit traffic. Additionally, the system correctly scales in during the session no. 345 due to the current demand dropped below the threshold, but then the values of the client demands increase above the allowable limit from session 355 onwards, requiring the system to scale-out again. Given that short-term network adjustments entail considerable overhead, introducing a client demand forecasting mechanism could enhance the system's ability to anticipate changes, reducing unnecessary adjustments and ensuring smoother performance under fluctuating conditions.

Proactive CU-UP scaling system

In Figure 5.8, the graphs of a CU-UP Horizontal Scaling System using a prediction mechanism under the same demand pattern as the previous two experiments are presented. As shown in Fig. 5.8a and 5.8b, the system can satisfy the requirements of both UEs in normal and peak periods with high demand conditions. The LSTM described in subsection 5.4.3 is able to predict with satisfactory accuracy, and the forecast values are also shown in Fig. 5.8a and 5.8b. Through the forecasting mechanism, the proposed system performs proactive scaling at session 211, compared to the case of reactive scaling system, which performs scaling-out at session 218. Due to the proactive scaling, the 5G network is able to satisfy user demands throughout its operation, since this approach prevents the severe performance degradation and packet loss observed in the real-time monitoring system, as scaling operations are performed in advance of demand spikes. The green and blue lines mark the moments where scaling-out and in occur respectively.

Another advantage of the proactive scaling system over the reactive one is the prevention of the ping-pong scaling effect, characterized by unnecessary scale-in at session 345 followed by scale-out at session 355. By effectively predicting the future increase in UE requirements,

the proposed system avoided these frequent adjustments of the CU-UP instances, which could otherwise introduce additional processing overhead and impact performance.

However, when the bandwidth values of the UEs are less than the threshold, the xApp triggers a scale-in system procedure, adhering to the desired operation. This behavior minimizes scaling overhead while ensuring sufficient capacity to accommodate fluctuating user demands. The forecasting mechanism thus enables efficient resource allocation and maintains high service quality, demonstrating its advantage over real-time monitoring alone.

5.6 Discussion & Limitations

Our findings highlight the effectiveness of integrating a proactive CU-UP scaling mechanism in 5G O-RAN systems, as a means of mitigating network congestion and successfully maintaining optimal performance. The proposed method was validated in terms of its effectiveness compared to reactive methods. Using ML, more specifically an LSTM-based traffic prediction model, the proposed proactive CU-UP scaling mechanism effectively anticipates network load variations, ensuring timely scaling decisions that prevent congestion of the CU-UP component. Compared to existing works limited to simulation environments, our real-world deployment and evaluation emphasize the applicability and feasibility of such mechanism integration in 5G commercial networks.

Given the time-sensitive nature of our proposed system, since decision-making happens on a real-time controller, greater emphasis could be placed on minimizing the ML inference time by exploring alternative models that could lead to measurable improvements in the system performance. Since our prediction implementation relies on a pre-trained model, the SMO could frequently retrain the model with new incoming data traffic to enhance adaptability and improve the system's accuracy.

Our proposed mechanism can also be extended to address similar network congestion bottlenecks, such as the UPF in the 5GCN or DUs in scenarios involving multiple RUs. However, an important aspect not discussed in this chapter is the placement of scaled CU-UP instances to minimize overall latency in a 5G network. The placement problem could be formulated as an optimization problem to determine the optimal placement strategy for the scaled CU-UP instances across the available nodes, aiming to maximize data traffic while minimizing latency constraints.

The proposed scheme has the potential to address other aspects impacting the network as well. For example, monitoring the load that each RAN component gets can be directly projected to the carbon footprint of the respective component. This in turn can be adjusted to a proactive scaling-in order to maintain the overall energy efficiency of the infrastructure. Similarly, proactive scaling can be applied as a method to mitigate vulnerabilities as they are detected in the network. For example, in case of identifying that the system is being used by malicious users, the users/slice that they belong can be transferred to a constrained CU-UP instance, so as to limit their overall impact to the system.

Nevertheless, the greatest limitation of our approach is the absence of the CU-UP Change procedure in our setup. This absence has been managed in our experimental procedure, in order to impact our results in the least possible manner, by redirecting the traffic of already connected UEs through instances of the CU-UP that already exist. This has been validated through our experimental results and was proven to have no impact on the scaling decision and the overall throughput that we manage to get, but merely impacts the seamless experience that a real user shall get over the system.

5.7 Conclusion

In this study, we have implemented a real-world 5G O-RAN network with a proactive CU-UP scaling mechanism that forecasts traffic demand using an LSTM model. Our system utilizes a real-world testbed environment in order to evaluate the effectiveness of the proposed system under dynamic network conditions. Based on the outcomes of our experiments, we validated that a proactive CU-UP scaling mechanism can enhance the system's ability to maintain QoS by performing CU-UP scaling before peak periods. We also highlight the importance of integrating ML models to forecast network load, as this integration minimizes packet loss, reduces unnecessary scaling overhead, and ensures smoother operation during high-demand periods. For future work, we aim to expand our approach by incorporating additional RAN components into the scaling mechanism and investigating optimization strategies for CU-UP placement to minimize latency and maximize throughput between the UE and the 5G network. As system utilization can directly affect the infrastructure footprint, scaling network functions can significantly impact overall network energy consumption. Therefore, in Chapter 6, we seek to further explore energy-efficient scaling strategies that balance system

performance and power consumption, which is becoming a key factor in making 5G networks more sustainable.

Chapter 6

Energy-Efficient Proactive CU-UP Scaling in 5G O-RAN: Vertical and Latency-Aware Horizontal Strategies

6.1 Introduction

Cellular networks have become a significant contributor to the environmental impact of the Digital Communication Technologies (DCT) sector, which accounts for the 2% of global greenhouse gas emissions as reported in [76].The RAN is responsible for over 70% of the total energy usage [18] of a 5G network, making energy efficiency a first-class design concern alongside performance towards 6G evolution. The cloud-native and disaggregated O-RAN architecture introduced in Chapter 2 provides the architectural foundation to address the high power consumption by deploying RAN components such as the CU-UP as VNFs orchestrated by Kubernetes, and by enabling the ML-driven control through xApps at the Near-RT RIC, it becomes possible to dynamically adapt resource provisioning in order to enhance the energy efficiency.

Chapter 5 highlights that the CU-UP is the primary bottleneck in the 5G RAN UP under high traffic conditions, and proposed a proactive horizontal scaling mechanism that successfully minimizes congestion and maintains QoS by deploying additional CU-UPs before demand peaks occurred. However, two fundamental questions were left unanswered. First, *what is the energy cost of these scaling decisions?* Performing horizontal scaling or maintaining overprovisioned resources consumes unnecessary power and without an energy-aware

scaling strategy, the system may consume more power in order to deal with the congestion. Second, *is horizontal scaling always the right tool?* An alternative strategy exists in the form of vertical scaling, where instead of spawning new instances, the CPU resources allocated to a single active CU-UP instance are dynamically adjusted and potentially follow a more energy-efficient approach. These two open questions converge into **RQ4**: **How can the different RAN User Plane scaling strategies be designed to jointly optimize performance, latency, and energy efficiency in disaggregated O-RAN deployments?**

In this chapter, we deploy a fully disaggregated 5G O-RAN network that supports the CUPS of the CU to CU-CP and CU-UP, alongside a near-RT RIC, and a service-based 5G CN, in a fully cloud-native environment orchestrated using Kubernetes (k8s). We propose two different energy-efficient and traffic-aware CU-UP scaling strategies implemented as xApps running on the near-RT RIC. The first xApp performs vertical scaling by dynamically adjusting the computational resources allocated to an active CU-UP instance to minimize 5G RAN power consumption, while the second implements latency-aware horizontal scaling by proactively instantiating or removing multiple CU-UP instances across the infrastructure to jointly reduce power consumption and minimize latency in the 5G O-RAN. Unlike simulation-based approaches, the proposed mechanisms are evaluated on a testbed using traffic traces sourced from a real-world dataset and power consumption measurements obtained via the open-source monitoring tools, thereby demonstrating their practical applicability under realistic operating conditions. This work aims to bridge the gap between O-RAN energy optimization theory and deployable practice.

The main contributions of this work are summarized as follows:

- Experimental analysis of the power consumption behavior of the CU-UP component under dynamic traffic conditions and resource provision.

- Design and implementation of energy-efficient traffic-aware CU-UP scaling mechanisms, supporting either vertical or latency-aware horizontal scaling, deployed as xApps at the near-RT RIC.

- Comparative evaluation of vertical and horizontal CU-UP scaling strategies, highlighting the trade-offs between energy efficiency and latency in practical O-RAN deployments.

6.2 Related Work

Energy efficiency has become a critical challenge in modern cellular networks, particularly as 5G and O-RAN architectures introduce greater flexibility, disaggregation, and cloud-based operation. The introduction of NFV and Cloud-RAN architectures in 5G brings the flexibility and resource pooling required for more adaptive and energy-efficient network operations. In this section, we review the relevant literature on scaling mechanisms in 5G networks, focusing on both the 5G CN and various RAN architectures such as monolithic gNB, Virtualized RAN (V-RAN), and O-RAN deployments, together with studies that introduce techniques for enhancing the energy efficiency of the 5G RAN.

In this context, VNF scaling has been widely discussed. For example, authors in [59] and [60] explore proactive VNF scaling driven by demand forecasts, leveraging online learning and optimization techniques. Their findings showcase reduced operational costs and improved service performance under dynamic conditions. Similarly, in [61], authors examine a scaling workflow based on predicted UE traffic, benchmarking ARIMA, XGBoost, and LSTM models on a real-world dataset that captures load behavior from 4G base stations. However, the approach is not applied to any specific 5G network function. Scaling in the 5G Core Network has been extensively explored in works such as [62] and [63], which focus on AMF scaling, while [65] examines the corresponding scaling mechanisms for the UPF. Regarding RAN scaling, [77] applies ML-based traffic prediction to trigger vertical scaling of the gNB and horizontal scaling of the UPF, offering one of the few practical approaches toward RAN-side elasticity. Likewise, authors in [70] use ML forecasting and an Integer Linear Programming (ILP) model to perform horizontal scaling of the CU, DU, and near-RT RIC. Nevertheless, the evaluation is limited to simulated scenarios. Our prior works [4] and [5] advance state-of-the-art on the RAN scaling by focusing on the underexplored CU-UP component, introducing a reactive horizontal scaling mechanism triggered by CPU resource usage [4] and a proactive horizontal scaling strategy based on ML-based demand prediction [5]. However, neither work considers the energy implications of the scaling decisions nor the impact on end-to-end latency.

Initially, research regarding energy savings in cellular networks was limited to sleep-mode base station approaches, where under-utilized cells were selectively switched off during low-traffic periods to reduce the power consumption as indicated in works [78], [79]. The same direction was also introduced in disaggregated 5G-RAN architectures, in which the

traditional base station power switching evolved into coordinated CU–DU–RU power management. In this setup, the primary focus was limited to implementing DU–RU sleep mode strategies [80]. Regarding O-RAN scenarios, authors in [81] propose an xApp responsible for dynamically controlling the RU sleep and wake-up states based on real-time traffic load. The overarching goal of this work is to reduce the RU power consumption while preserving QoS guarantees, utilizing real RU equipment.

RAN Energy Efficiency for 5G networks has been studied with respect to the consumption of the radio components. Authors in [82] follow a radio resource allocation strategy, optimizing transmission power, user scheduling, and bandwidth allocation in order to enhance the energy efficiency at the radio interface. The work is limited to simulation scenarios. Another approach described in [83] enhances energy efficiency in multi-user MIMO systems by jointly optimizing the number of antennas, active users, and transmit power under a realistic circuit power model.

Recently, numerous studies such as [84] develop and standardized the energy measurement methodologies of the O-RAN. In this work, we aim to extend this foundation by modeling the power consumption of the CU-UP component and design scaling strategies for real-time network control. Similarly to our approach, paper [85] performs a comprehensive analysis of power consumption of the O-RAN network components, modeling the individual energy contributions of the O-RU, O-DU, and O-CU components under different traffic conditions and functional splits. A similar approach was also followed in [86], in which the energy consumption behavior of monolithic and distributed RAN architectures was collected and modeled using open-source monitoring tools like Kepler and Scaphandre in Kubernetes environments. Pawar et al. illustrated in [87] profile the energy consumption of an OpenAir-Interface (OAI) based Cloud-RAN (C-RAN) implementation and propose a traffic-aware compute-resource tuning scheme that adapts the CPU frequency of the host running the CU, DU and RU functions. The authors simulate the proposed scheme using the *ns3* simulator, demonstrating up to 40% power savings without degrading throughput.

Additionally, Hojeij et al. [88] propose a joint ILP- and Graph Neural Network (GNN) -based framework for energy-efficient UE association and CU/DU placement in disaggregated O-RAN environments, achieving large theoretical savings using analytical power models. However, their results solely rely on simulation and formula-based energy estimation, without real testbed measurements. In our prior work [9], we examined the energy consump-

tion of the CU component in a typical 5G O-RAN deployment in the OneLab testbed and proposed an rApp that performs horizontal CU scaling based on a predefined data-volume sequence. However, that proof-of-concept assumes that the requested traffic volumes and their transmission duration are known in advance, and minimizes the total energy to deliver them by trading CU count against completion time. In contrast, this work targets the CU-UP component, uses ML-based forecasting to act on unknown future demand, and, via its scaling decisions, minimizes the instantaneous power needed to sustain the predicted throughput, making it applicable to operational O-RAN deployments.

A review of the existing literature reveals that most real-world, testbed-driven contributions on energy efficiency in O-RAN and 5G RAN infrastructures predominantly target the reduction of power consumption at the RU level ([81], [80], [82], [83]), with additional efforts focusing on the DU ([80], [87]) and, to a lesser extent, the CU ([9], [88]). In this work, we combine and expand upon the existing methods discussed in the literature by focusing on the CU-UP component and proposing two proactive scaling strategies driven by UE's traffic forecasting: a vertical and a latency-aware horizontal, having as objective the reduction of the CU-UP power-consumption within a 5G O-RAN network. We then compare the performance of the two strategies and validate their effectiveness using monitoring tools.

6.3 CU-UP Power Consumption Modelling

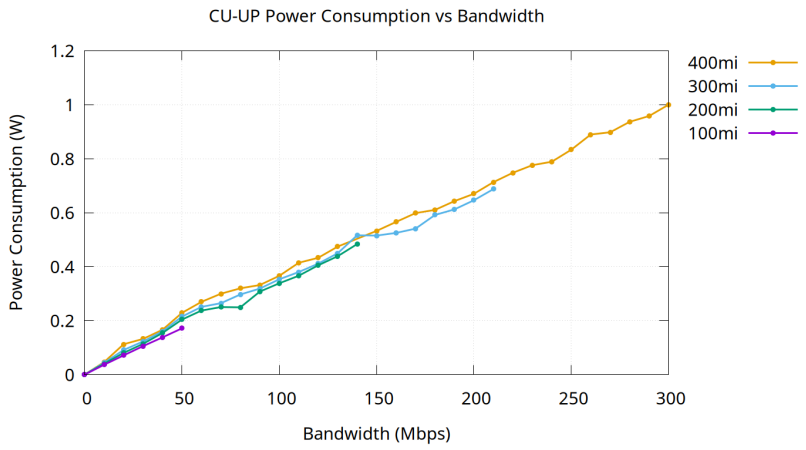


Figure 6.1: Impact of CPU Resources on CU-UP Power Consumption vs Bandwidth

In contrast to prior work that relies on analytical power models [88], we follow an experimental based approach to derive a CU-UP power consumption model based on real-world

measurements collected under varying traffic loads. In O-RAN, the CU-UP is deployed as a VNF. Following the same methodology described in papers [86, 89], we utilize Scaphandre [90], an open-source energy monitor tool to monitor the power consumption of the CU-UP

Table 6.1: Comparison of state-of-the-art with our approach.

Works	**Approach**	**Evaluation environment**	**Scaling**	**Energy Efficient**
[62], [63]	AMF scaling in the 5G Core	testbed environment	Yes	No
[65]	UPF auto-scaling	testbed environment	Yes	No
[77]	ML-based vertical scaling of gNB and horizontal scaling of UPF	simulated environment	Yes	No
[70]	ILP + ML for scaling CU, DU and Near-RT RIC	simulated environment	Yes	No
[4]	Reactive CU-UP horizontal scaling	testbed environment	Yes	No
[5]	Proactive ML-based CU-UP horizontal scaling	testbed environment	Yes	No
[81]	xApp-driven dynamic RU sleep/wake control	testbed environment	No	Yes
[80]	Coordinated CU–DU–RU sleep-mode energy saving	testbed environment	No	Yes
[87]	Traffic-aware CPU frequency tuning for C-RAN	testbed environment	No	Yes
[9]	rApp for horizontal CU scaling via predefined data-volume data traffic	testbed environment	Yes	Yes
[88]	Energy-Efficient Placement of O-RAN components	simulation	No	Yes
This work	**Proactive Vertical or Horizontal CU-UP scaling based on traffic prediction**	**testbed environment**	**Yes**	**Yes**

under varying traffic conditions. Scaphandre relies on the RAPL (Running Average Power Limit) sensor and the Linux powercap module to capture the power consumption of the entire host over time. This tool monitors the power usage of an individual process by first collecting the total CPU power consumption and then estimating each share according to the amount of time the process utilized the CPU. In our case, the CU-UP VNF power consumption is computed by aggregating the consumption of all processes belonging to the CU-UP container and exporting the results via the Prometheus monitoring tool.

Inspired by paper from Pawar et al. [87], in which the CPU frequency was adjusted and directly regulated the energy consumption of C-RAN functions, our approach translates this idea into a vertical-scaling framework where we restrict the available CPU resources of the CU-UP.

In our testbed-based experiments, the CU-UP was provisioned with varying CPU resource levels (from 100 millicores (mi) to 400mi) for each profile, thereby affecting the traffic volume it can effectively handle. The requested served traffic was increased by 10 Mbps in each step, from 0 to 300 Mbps, which is the maximum throughput our 5G setup can handle, as described in 6.5.1. Each experiment was repeated 24 times to ensure robustness and for each step we removed the 4 outliers (maximum and minimum values) and reported the median power measurements. Through these tests, we also identified the maximum throughput each CU-UP provisioning level could serve: 100mi sustains up to 50 Mbps, 200mi up to 130 Mbps, 300mi up to 210 Mbps, and 400mi up to 300 Mbps. Requesting traffic beyond the profile's maximum threshold (e.g., exceeding 50 Mbps on the 100mi profile) degrades throughput, as the CU-UP cannot sustain the requested rate within the expected time, increasing serving time. These thresholds serve as a foundation for analyzing demand-aware resource provisioning. It is important to highlight that the reported power measurements correspond to a specific CPU model deployed on our CU-UP nodes and described in Section 6.5.2. Other CPU models/architectures may introduce slight increases or decreases in the absolute power consumption values.

Observation 1. *A key observation is that, when the CU-UP is provisioned with the minimum amount of resources, but sufficient enough to meet the UE demands, it consistently consumes less power compared to deployments with higher resource allocations.*

For instance, at 50 Mbps, all provisioning levels (100mi to 400mi) are capable of handling the traffic, yet the 100mi configuration consistently exhibits the lowest power consumption.

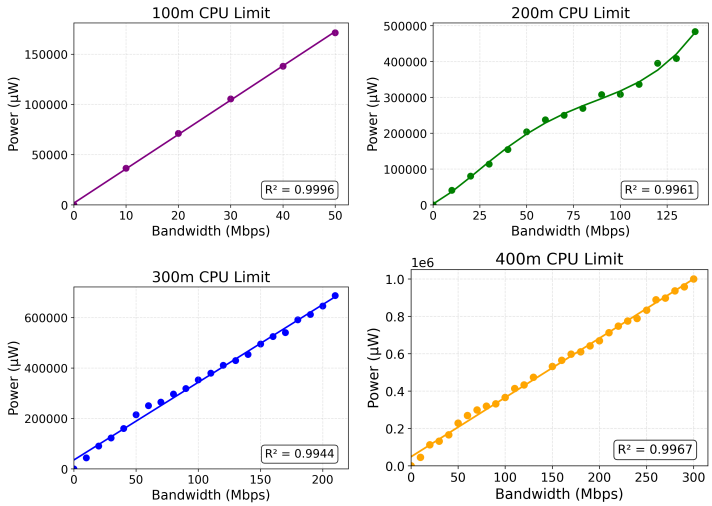


Figure 6.2: Regression Models for CU-UP power consumption

This behavior is observed across all tested bandwidth levels and occurs because lower CPU allocations restrict the CU-UP to a smaller share of the CPU time, as enforced by the Linux Completely Fair Scheduler (CFS) bandwidth controller [91], reducing the number of active processor cycles and consequently lowering its overall power consumption. Since each traffic session has a fixed duration ΔT and the minimum sufficient profile sustains the demanded throughput without degradation, this power reduction directly translates to lower energy consumption ($E = P \times \Delta T$). However, as discussed above, this equivalence holds only when the CU-UP operates within its capacity bounds, as exceeding a profile's threshold degrades throughput and violates the QoS requirements of the served UEs.

Our study resulted in the power consumption behavior illustrated in Figure 6.1. We created 4 different Power consumption estimator models across the CPU provisioning profiles, as illustrated in Figure 6.2. For the 100 mi, 300 mi, and 400 mi CPU limits, the empirical measurements exhibited an almost linear pattern, and thus we fitted standard linear regressions to estimate the corresponding power-bandwidth functions. In contrast, the 200 mi profile exhibited a nonlinear relationship, so we used a third-order regression.

6.4 System Architecture

This work studies the application of scaling techniques at the CU-UP O-RAN's component to minimize the operational power consumption of a fully disaggregated 5G O-RAN. In general, scaling follows two different strategies: vertical scaling, where the resources of

an already existing instance are decreased or increased (scale down or up accordingly), and horizontal scaling, in which additional instances are deployed or undeployed (scale out or in accordingly) across the infrastructure. The proposed system architecture, as seen in Figure 6.3, follows a fully disaggregated, cloud-native 5G O-RAN where the CU-UP is dynamically scaled vertically or horizontally by AI-driven xApps at the near-RT RIC, with the scaling mode determined by the instantiated xApp. Focusing more on the 5G-RAN, it follows the 3GPP Option-2 functional split, with CU-CP, CU-UP, and DU deployed as VNFs and being interconnected via standardized interfaces (E1 and F1), while the CU-UP provides the user-plane data path between the DUs and the 5G CN over N3. To enable proactive scaling, an ML model is used to forecast future UE throughput demands [5]. These forecasting model is trained by the Service Management and Orchestration (SMO) framework and provided to the xApps, which make scaling decisions based on the predicted traffic and real-time KPIs collected over the E2 interface.

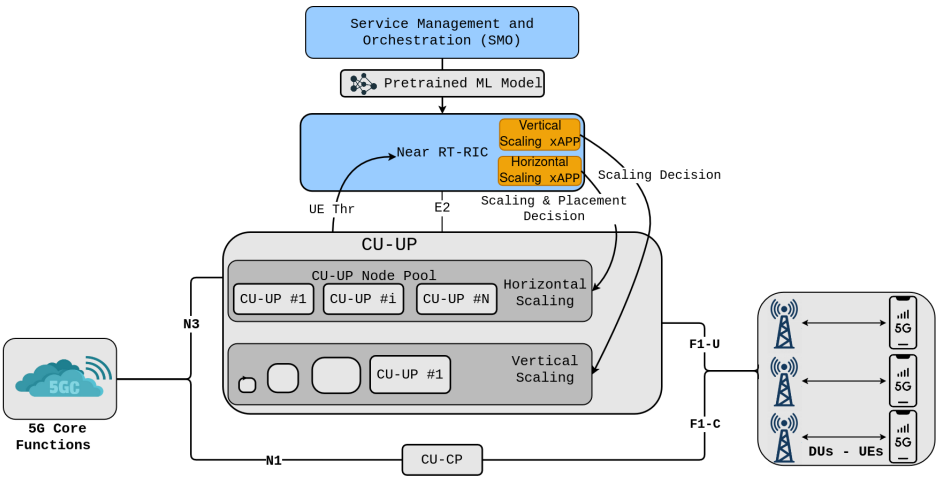


Figure 6.3: 5G O RAN Architecture Supporting CU UP Vertical or Horizontal Scaling Through Dedicated xApps

6.4.1 Machine Learning Model

Forecasting the future data network demands is a fundamental requirement for efficient network resource management. In our approach, this problem is formulated as a time-series forecasting task, where past traffic patterns are used to predict future demand.

Dataset

The dataset used to train the ML model is described in detail in Section 2.7. Following the same approach as in Chapters 4 and 5, we retain only the DL bitrate feature for throughput forecasting, and the first three day-snapshots are used for training. During preprocessing, min-max scaling was applied within the operating range of 0–130 Mbps per client, as defined by the CU-UP capacity constraints of this chapter's experimental setup (Section 6.5), followed by spline interpolation to reduce data noise.

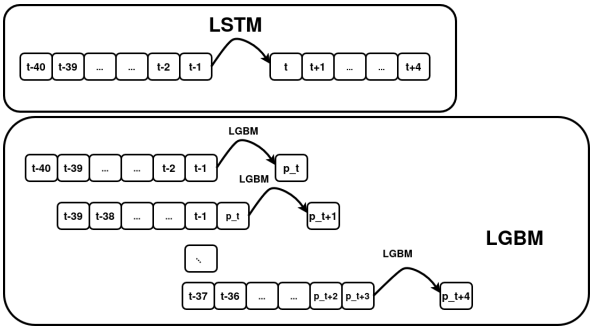


Figure 6.4: ML models prediction architecture

Forecasting Methodology

To enable proactive operation, a multi-step forecasting is required. By predicting the future values, the system gains the proactive lead time needed to initiate scaling, guaranteeing resources are ready before the predicted high-demand period arrives. By performing multi-step forecasting, the system prevents the ping-pong effect where the system continuously changes state in response to transient traffic fluctuations. Avoiding such unnecessary state changes reduces redundant control and network operations of the 5G O-RAN and potential service downgrade. For the purpose of this work, a 5-step ahead prediction horizon is sufficient for CU-UP scaling decisions.

Two different multi-step prediction methods exist: 1) a method that follows a recursive prediction and 2) a method that follows a direct one. The recursive methodology is based on predicting one time step and feeding it back to the model as an input for the subsequent prediction. On the contrary, the direct one implies that a multi-output approach is used in which the model is trained to directly predict all the future values simultaneously.

LSTM

The first ML model we evaluated was a the LSTM used in previous Chapters in Sections 4.3.4 and 5.4.3, which was trained using the dataset as defined in Section 2.7. Comparing these two approaches, the direct one achieved better performance. The direct model consists of two stacked LSTM layers with 256 and 128 neurons, followed by a dense output layer, trained to predict the next 5 time-series values given an input sequence of 40 past values. We utilized the keras-tuner in order to select the optimal configuration based on the training dataset and the required prediction horizon. Training was performed for 15 epochs using the Adam optimizer with a learning rate of 0.0015, achieving a mean absolute percentage error MAPE of 4.9% with an average inference time of 28.76ms.

LightGBM

The second ML model that we evaluated in our study is the Light Gradient Boosting Machine (LightGBM), a decision-tree–based ensemble method especially designed for extremely low inference latency, a really important feature for our real-time scaling xApp being deployed at the nearRT-RIC. In contrast to recurrent neural architectures, LightGBM does not explicitly model sequential dynamics, but instead it learns nonlinear relationships over fixed-length input windows. Since LightGBM is a regressor, there are two primary methods for performing multi–time step forecasting. The first approach employs multi-step regressors, in which a separate model (one regressor per forecast time-step) is trained to predict all future horizons. The second approach is the traditional recursive strategy, where a single-step model is iteratively applied, using each newly generated prediction as an input to forecast subsequent time steps.

Evaluating both methods, the LGBM with the recursive strategy achieved better performance, achieving a MAPE of 4.7% and an average inference time of 1.2 ms. This model uses a 40-step input window and is configured with 64 leaves, a 0.05 learning rate, feature and bagging fractions of 0.9, and 400 boosting rounds to ensure stable learning.

Putting everything together, Figure 6.4 illustrates the prediction architectures of the evaluated ML models, highlighting the direct multi-step forecasting approach used by the LSTM and the recursive forecasting strategy adopted by the LightGBM model. Among these two ML models, The LightGBM model was selected as the forecasting model for the deployment in the xApp, since it offers lower MAPE and substantially faster inference times compared

to the LSTM.

The lowest inference time is crucial for real-time predictions at the near-RT RIC, where inference latency constraints are stringent. It is important to note, however, that our system currently requires only short-horizon forecasting, limited to only five future steps. This short-window setting inherently favors tree-based regressors such as LightGBM, which can efficiently capture nonlinear dependencies over fixed-length inputs. For systems that require longer prediction horizons, LSTM-based architectures are expected to become more advantageous due to their ability to model long-range temporal dependencies and sequence dynamics more effectively.

6.4.2 Vertical Scaling

In this subsection, we will describe the algorithm utilized by the first xApp that implements vertical scaling of the CU-UP instances in a 5G O-RAN infrastructure. Our target is to minimize the energy consumption of a 5G system required to satisfy the network demands of the connected UEs, via proactive scaling.

This strategy follows the **Observation 1**, which shows that allocating the minimum resources required by the CU-UP to serve the traffic consistently reduces power consumption compared to overprovisioning. Building on this foundation, we design a vertical scaling xApp that proactively scale-up when the forecast traffic exceeds current provisioning, and reactively scales-down only when real-time measurements indicate persistent under-utilization. The Algorithm 4, uses the LGBM model defined in Section 6.4.1 to predict per-UE throughput over a 5-step future time horizon. It subsequently determines the required CU-UP profile based on experimental thresholds taken from real-world measurements as described in Section 6.3. These thresholds correspond to provisioning levels of 100–400mi CPU that were experimentally validated in Section 6.3, where each level sustains a known maximum throughput with minimal power overhead. Importantly, the xApp integrates the energy model into the scaling logic by always selecting the profile that minimizes the energy consumption to serve the predicted demand, ensuring energy-efficient operation.

6.4.3 Horizontal Scaling

In this subsection, we focus on the problem of horizontally scaling CU-UP instances and their placement across candidate nodes in a 5G O-RAN infrastructure. Compared to the

Algorithm 4 xApp Vertical Proactive Scaling

Require: Window Size W, Forecast Interval T, Prediction Horizon P

Require: Pretrained LGBM Model, Profile Thresholds $\{100, 200, 300, 400\}$ Mi

Initialize: Sliding window $S[UE]$ with W historical DL throughput values per UE
Initialize: Real-time throughput $R[UE] \leftarrow 0$ for all UEs
Initialize: Load pretrained LGBM model
Initialize: Set $\text{Profile}_{\text{current}} \leftarrow 100$
Initialize: Connect to RT-RIC and subscribe to E2SM-KPM
while true **do**
    Fetch current DL throughput $R[UE]$ for each UE
    Update $S[UE]$ by appending $R[UE]$ and removing the oldest entry
    if time mod $T = 0$ **then**
        $\text{ThrPr}_{\text{total}} \leftarrow 0$
        for all UE (in parallel) **do**
            $\text{ThrPr}[UE] \leftarrow \text{LGBM}(S[UE])$
            $\text{ThrPr}_{\max}[UE] \leftarrow \max(\text{ThrPr}[UE])$
            $\text{ThrPr}_{\text{total}} \leftarrow \text{ThrPr}_{\text{total}} + \text{ThrPr}_{\max}[UE]$
        end for
        $\text{Profile}_{\text{target}} \leftarrow \texttt{ProfileSelector}(\text{ThrPr}_{\text{total}})$
        if $\text{Profile}_{\text{target}} > \text{Profile}_{\text{current}}$ **then**
            Proactive Scale-Up: Set CU-UP profile to $\text{Profile}_{\text{target}}$
            $\text{Profile}_{\text{current}} \leftarrow \text{Profile}_{\text{target}}$
        end if
    else
        $R_{\text{total}} \leftarrow \sum_{\text{UE}} R[UE]$
        $\text{Profile}_{\text{actual}} \leftarrow \texttt{ProfileSelector}(R_{\text{total}})$
        if $\text{Profile}_{\text{actual}} < \text{Profile}_{\text{current}}$ **then**
            Reactive Scale-Down: Set CU-UP profile to $\text{Profile}_{\text{actual}}$
            $\text{Profile}_{\text{current}} \leftarrow \text{Profile}_{\text{actual}}$
        end if
    end if
end while

vertical scaling that focuses solely on power consumption minimization, horizontal scaling can also minimize latency in the RAN subsystem, since the scaled-out instances can be placed across multiple nodes in the 5G O-RAN, resulting in different latency levels across other 5G components (e.g DU) in the infrastructure, while still reducing energy consumption.

Problem Definition

We describe the problem of CU-UP scaling and instance placement in a 5G cloud-based RAN infrastructure as follows: Initially, we are given a set of candidate nodes for deploying CU-UP instances, alongside an additional set with the nodes where DUs have been deployed. We also consider the link latency between these candidate nodes and the nodes hosting the deployed DUs. This optimization problem is formulated after the xApp predicts the future traffic demands of each DU, obtained by aggregating the predicted demands of its connected UEs. We follow this strategy since our goal is not only to reduce the overall energy consumption but also to minimize the latency between each DU and its assigned CU-UP via the optimal placement. Furthermore, the capacity of each CU-UP is specified as a constraint. Since our goal also includes reducing the energy consumption of the RAN, we take into account the energy required by each CU-UP to serve a specific amount of data plane traffic. To study this, we assume that CPU power usage is the main contributor to the energy consumption of the CU-UP, excluding other components (e.g., memory, fans). This assumption aligns with the ETSI TS 28.310 standard [92], which supports that the energy consumption of a VNF instance is determined by its proportion of the total NFV node consumption based on its CPU usage.

Problem Formulation

The definition of the proposed ILP formulation is presented as follows:

Sets:

- N : Set of candidate nodes to deploy CU-UP .
- M : Set of nodes DUs are deployed.

Parameters:

- lat_{ij} : Latency between DU node i to CU-UP node j.
- w_i : Traffic demand of DU i.
- cap_j : Capacity of CU-UP j.
- B : A large constant chosen such that energy minimization is lexicographically prioritized over latency minimization (e.g., $B > N \cdot \max_{i,j} lat_{ij}$).

- $E_j(T_j)$: Energy consumption function of CU-UP j, where T_j is the total throughput served by CU-UP j during a fixed time interval Δt:

$$E_j(T_j) = \begin{cases} 0, & \text{if } T_j = 0 \\ P(T_j) \cdot \Delta t, & \text{if } T_j > 0 \end{cases}$$

where $P(T_j)$ denotes the power consumption function of CU-UP j derived in Section 6.3.

Decision Variables:

- $x_{ij} \in \{0, 1\}$, where:

$$x_{ij} = \begin{cases} 1, & \text{if DU } i \text{ is assigned to CU-UP } j \\ 0, & \text{otherwise} \end{cases}$$

- $y_{jl} \in \{0, 1\}$, where:

$$y_{jl} = \begin{cases} 1, & \text{if CU-UP } j \text{ is activated} \\ 0, & \text{otherwise} \end{cases}$$

- T_j : Total throughput handled by CU-UP j, defined as:

$$T_j = \sum_{i=1}^{N} w_i x_{ij}$$

Objective:

$$\min \quad B \sum_{j=1}^{M} E_j(T_j) + \sum_{i=1}^{N} \sum_{j=1}^{M} (lat_{ij} \cdot x_{ij})$$

The above objective function aims to perform a joint optimization of energy consumption and latency for CU-UP deployment and DU assignment in a 5G O-RAN network and is subject to the following constraints:

Unique Assignment of DUs:

$$\sum_{j=1}^{M} x_{ij} = 1, \quad \forall i \in \{1, \ldots, N\} \tag{6.1}$$

Constraint 1 prevents unassigned DUs, ensures efficient resource utilization, and follows the 3GPP rule that each DU is assigned to exactly one CU-UP.

Capacity Limits:

$$T_j \leq cap_j y_j, \quad \forall j \in \{1, \ldots, M\} \tag{6.2}$$

Constraint 6.2 ensures that a CU-UP can only handle traffic if it is active and within its capacity, preserving the power-to-energy equivalence established in Observation 1.

Coupling of Assignment and Activation:

$$x_{ij} \leq y_j, \quad \forall i \in \{1, \ldots, N\}, j \in \{1, \ldots, M\} \tag{6.3}$$

Constraint 6.3 ensures that a DU can only be assigned to a CU-UP if that CU-UP is active, preventing assignments to inactive CU-UPs and ensuring proper 5G RAN functionality.

This optimization problem can be viewed as a variant of the Facility Location Problem (FLP) as described in [93]. The FLP aims to determine the optimal locations for facilities to meet customer requirements at the lowest possible cost. In the CU-UP placement problem, each CU-UP node acts as a facility, and the DUs serve as clients. The objective is to determine an optimal set of CU-UPs to activate and assign DUs while minimizing the combined cost of latency and energy consumption. Since FLP is known to be NP-hard [93], and our problem generalizes it by introducing additional constraints on latency and energy efficiency, it follows that the proposed CU-UP placement problem is also NP-hard. To address the high computational complexity of the proposed optimization problem in large-scale infrastructures, we introduce a heuristic, shown in Algorithm 5. This greedy algorithm follows the same principles, aiming to minimize the total energy consumption of CU-UP instances while achieving the minimum DU to CU-UP latency within a 5G O-RAN architecture. The main idea is to identify the high-demand DUs first and map them to CU-UPs in a way that minimizes additional energy usage and latency. Since the algorithm only assigns traffic to a CU-UP when its remaining capacity is sufficient to serve the demand, the serving time is never extended, and therefore minimizing power directly translates to minimizing energy as established in Observation 1.

Initially, the heuristic initializes the available power budget of each CU-UP based on the selected resource profile, using the power models described in Section 6.3. Each profile defines both the maximum traffic that a CU-UP instance can sustain and the corresponding maximum power it is expected to consume. At every predefined control interval, the xApp implementing the latency-aware horizontal scaling heuristic first predicts the aggregate traffic demand of each DU using the LGBM model and then sorts the DUs in descending order of

predicted load. In the next stage, the heuristic iterates over the DU list sorted by demand, computes the power required to serve each DU, and attempts to assign it to an active CU-UP with sufficient remaining power budget, while minimizing the midhaul latency. If no such CU-UP exists, a new CU-UP instance is activated and placed on the node that offers the minimum DU to CU-UP latency. After each assignment, the remaining power budget of the selected CU-UP is updated to account for the additional load. This greedy strategy jointly minimizes the number of active CU-UP instances, constrains total power consumption, and favors low-latency DU–CU-UP associations.

In order to explore the applicability of the proposed heuristic, it is important to analyze its time complexity. During the preprocessing phase, the most computationally intensive task is sorting the DUs by demand, which has a complexity of $O(Nlog(N))$, N being the number of DUs. It is important to note that in our implementation, the predicted traffic demand of each DU is obtained by aggregating the predicted demands of its connected UEs. This design was chosen since the 5G community is lacking publicly available DU-level traffic datasets. Moreover, the aforementioned forecasting mechanism, as seen in Subsection 6.4.1, operates at the UE level, and DU demand is derived by summing the corresponding UE predictions. In a real-world deployment, DU demand could be predicted without traversing the individual UEs, resulting in significantly lower computational overhead. For this reason, the computational complexity of UE-level prediction, which would be $O(N \cdot U_{Ni})$ (U_{Ni} is the number of UEs served by each DU Ni), is not included in the complexity analysis of the proposed heuristic.

The rest of the initialization phase involves the setup of the energy consumed for all CU-UPs. This complexity is $O(M)$, while initializing the assignment array for all DUs is $O(N)$. During the main loop, responsible for the DU assignment, the heuristic iterates over each of the N DUs. During the first stage, it evaluates the currently active CU-UPs based on available power and latency. In the worst case, the active set can contain up to M CU-UPs, resulting in a cost of $O(M)$ per DU. If no active CU-UP is suitable, the algorithm proceeds to search among the inactive CU-UPs to construct a list of feasible candidates, which again takes $O(M)$ time.

Algorithm 5 xApp Latency-Aware Horizontal Proactive Scaling

Require: `lat`$[i][j]$: Latency from DU i to CU-UP j

`UEset`$[i]$: Set of UEs connected to DU i

`LGBMpredict`(ue): ML model for UE throughput prediction

`powerLimits`$[j]$: Max power budget of CU-UP j (W)

`powerModel`(d): Power required to serve demand d (W)

T: Control interval

Initialize: Connect to RT-RIC and subscribe to E2SM-KPM

$N \leftarrow$ number of DUs

$M \leftarrow$ number of CU-UPs

while true **do**

  Wait until next T-second control interval

  ▷ A) Predict future per-DU traffic

  for $i = 1$ to N **do**

    $demands[i] \leftarrow \sum_{ue \in UEset[i]}$ `LGBMpredict`(ue)

  end for

  ▷ B) Sort DUs by descending predicted demand

  $duOrder \leftarrow$ DUs sorted by $demands[i]$ (descending)

  ▷ C) Reset CU-UP power budgets

  for $j = 1$ to M **do**

    $remainingPower[j] \leftarrow powerLimits[j]$

  end for

  $activeCUs \leftarrow \emptyset$

  $assignment[i] \leftarrow -1$ for all DUs

  ▷ D) Greedy power–latency placement

  for each du in $duOrder$ **do**

    $needed \leftarrow demands[du]$

    $powerNeeded \leftarrow$ `powerModel`$(needed)$

    $bestCU \leftarrow$ None

    $bestLat \leftarrow \infty$

    ▷ Try active CU-UPs first

    for each cu in $activeCUs$ **do**

      if $remainingPower[cu] \geq powerNeeded$ **and** $lat[du][cu] < bestLat$ **then**

        $bestCU \leftarrow cu$

        $bestLat \leftarrow lat[du][cu]$

      end if

    end for

    ▷ If needed, activate a new CU-UP

    if $bestCU =$ None **then**

      $feasible \leftarrow \{c \notin activeCUs \mid remainingPower[c] \geq powerNeeded\}$

      $bestCU \leftarrow \arg\min_{c \in feasible} lat[du][c]$

      $activeCUs \leftarrow activeCUs \cup \{bestCU\}$

    end if

    $assignment[du] \leftarrow bestCU$

    $remainingPower[bestCU] \leftarrow remainingPower[bestCU] - powerNeeded$

  end for

  ▷ E) Apply scaling actions

  Deploy CU-UPs in $activeCUs$

  Reassign DUs according to $assignment$

end while

During the last stage, only assignments and updates to values are happening, requiring only constant time $O(1)$. So in the worst-case scenario, the complexity of the main loop is $O(N \cdot M)$. The overall complexity by combining both phases is $O(N \cdot M)$.

6.5 Experiments

We conducted extensive experiments to demonstrate the viability of the proposed solution. The evaluation was carried out on the NITOS testbed, the Greek node of SLICES-RI [27], located at the University of Thessaly.

6.5.1 5G Network Configuration and Setup

Our experimental topology consists of a cloud-based 5GCN & O-RAN network deployed across nine nodes (Node 1 - Node 9), utilizing the Kubernetes orchestration tool. All the 5G components use the OAI software suite, which provides a software-based, network function virtualized implementation of the 5GCN, RAN, and RIC. We utilized a basic 5GCN architecture which is decomposed in the following VNFs: AUSF, UDM, UDR, NRF, AMF, and SMF. Regarding the User Plane (UP) of the 5GCN, two software-based UPFs were utilized, in order to have the CU-UP as the only congestion point in the UP of the 5G CN & RAN paths. Regarding the 5G RAN, we use the fully disaggregated architecture where we support the 3GPP Option 2 split [12], which defines the division of functionalities between the CU, responsible for the PDCP, RRC, and SDAP layers, and the DU, which executes the RLC, MAC, and PHY layers, via the introduction of the F1 interface as specified in the 3GPP specifications TS 38.470–38.473 [94] and mentioned in Section 2.2.2.

Furthermore, we consider the logical Control/User-Plane separation of the CU into two components: the CU Control Plane (CU-CP) and the CU User Plane (CU-UP), interconnected via the E1 interface as specified in 3GPP TS 38.463 [95].

Regarding the deployment in the physical infrastructure, Node 1 was explicitly dedicated to host all the CP VNFs for the 5GCN and Node 9 is dedicated to the corresponding CU-CP component of the 5G RAN, together with the OAI-FlexRIC, the near-RT RIC. The Near-RT RIC plays a crucial role in orchestrating and optimizing the functions of RAN components by continuously collecting Key Performance Indicators (KPIs) and Key Performance Measurements (KPMs) from the E2 nodes. Following the methodology used in Chapter 5, we use the

same KPI *DRB.UEThpDl*, defined in the 3GPP TS 28.552 standard [72], which represents the downlink bitrate per UE measured at the CU-UP's DRB queues. Two nodes (Nodes 2 and 3) were used to deploy two separate UPF instances. Three additional nodes (Nodes 4, 5, and 6) were selected as candidate locations for deploying the CU-UP components.

In this deployment, we utilize two DUs and two UEs to increase the overall network capacity and validate multi-user performance. The connection between each UE and DU is not established over a real-world radio device but instead, we employ an RF simulator that emulates the behavior of a physical Radio Frequency (RF) board. To approximate real-world channel conditions, we rely on a model with Additive White Gaussian Noise (AWGN). The use of an RF simulator does not compromise the real-world applicability of our evaluation since the proposed system model does not rely on radio-specific KPIs, but solely on the UE throughput KPIs.

We test two different xApps: one responsible for Vertical Scaling and one for latency-aware Horizontal Scaling. Both xApps rely on a predefined mapping of the UE–DU assignment and the CU-UP–to–DU latency matrix, as defined in Table 6.3. Both xApps subscribe to E2SM-KPM to obtain the reported UE downlink throughput measurements, and they also employ the same LGBM model to predict the future demands of the UEs. Additionally, to accurately measure the energy impact of the scaling decisions, real-time power consumption of the CU-UP instances was continuously monitored using Scaphandre and the Prometheus exporter following the methodology described in Section 6.3.

Midhaul and Backhaul Latency Model According to [96], the midhaul (CU-UP-to-DU) latency of a 5G network is bounded by 10 ms. Based on these constraints, we configure the inter-node RTT values (assuming that RTT is equal to the two-way latency) of the midhaul segment into three representative operating classes: low latency (1.5 ms), medium latency (5 ms), and high latency (8.5 ms). For the backhaul (CU-UP-to-UPF), we adopt a fixed RTT of 5 ms, reflecting a typical non–latency transport configuration of the N3 interface. These RTT values were emulated in our testbed environment using the *netem* tool. These RTT values correspond to typical country-wide operator deployments, such as those observed in Greece. For instance, empirical measurements show an RTT of approximately 8 ms between the University of Thessaly (Volos, Greece) and National Technical University of Athens (Athens, Greece) servers, reflecting realistic nation-wide latencies.

Putting everything together, the overall experimental setup of the deployed architecture,

Table 6.2: Node Assignments and Specifications

Node	Components	CPU	Cores	RAM
1	NRF, UDR, UDM, AUSF, AMF, SMF	Intel Core Processor (Skylake, IBRS)	4	8GB
2,3	UPF (Slice: oai & ims)	Intel Core Processor (Skylake, IBRS)	4	8GB
4,5,6	CU-UPs	Intel Core Processor (Skylake, IBRS)	4	8GB
7	DU, UE (Pair 1)	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz	6	16GB
8	DU, UE (Pair 2)	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz	6	16GB
9	CU-CP, FlexRIC	Intel Core Processor (Skylake, IBRS)	4	8GB

Table 6.3: RTT Latency Between CU-UP and DU Nodes

CU-UP Node	DU Node	RTT (ms)
4	7	1.5
4	8	8.5
5	7	5
5	8	5
6	7	8.5
6	8	1.5

is shown in Figure 6.5.

6.5.2 Real-World Traffic Generation

To evaluate the proposed xApps and their power consumption, we employed two day-snapshots of the dataset as presented in Section 2.7 that were not included in the training dataset, which were concatenated and replicated to extend the trace and increase the system load. The generated traffic was carried out using *iperf* sourced from the N6 interface of the UPF and the UEs, with each value of the dataset replicated for a 5-second traffic session, im-

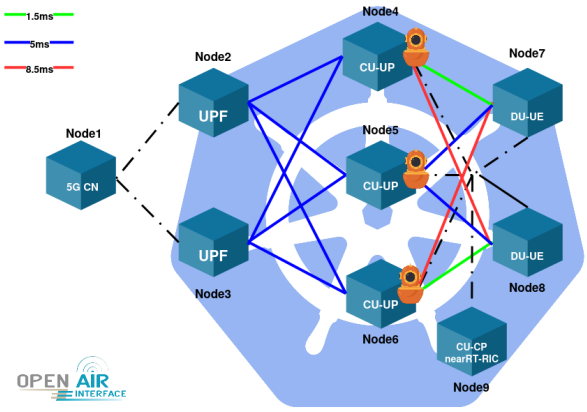


Figure 6.5: Experimental Architecture

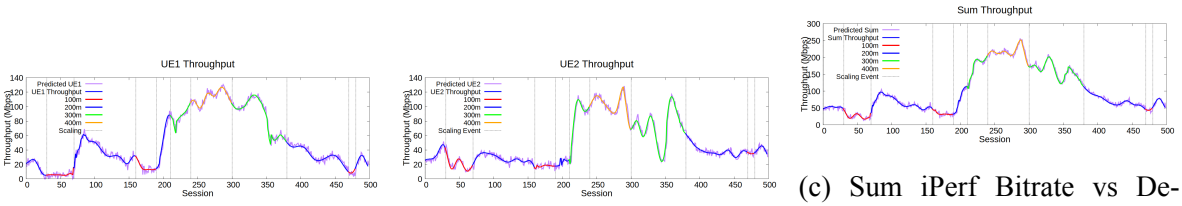
plementing a DL traffic scenario. We apply the same traffic pattern across all scenarios that implement the baseline system, the xApp with vertical scaling, and the xApp with horizontal scaling in order to ensure a fair and consistent comparison, guaranteeing that any observed differences are solely due to the applied scaling mechanisms. The baseline system is the strategy that is followed today by the industry following monolithic deployments where resources are allocated based on peak load conditions.

6.5.3 CU-UP scaling process

Performing scaling of the 5G VNFs in real-world cellular networks is considerably more complex compared to the cloud computing field. Vertical scaling of the CU-UP is relatively simpler compared to the horizontal scaling. Vertical scaling is based on a single CU-UP deployed instance with a unique ID, where only the CPU's resource allocation is adjusted according to the xApp's decisions. Thus, scale-up is achieved by allocating additional CPU resources to the CU-UP, while scale-down corresponds to reducing its CPU allocation resources. In this context, restarting the CU-UP deployment with the updated CPU resource configuration and the according UEs that are served from this is sufficient to return the 5G-RAN into a steady-state operation. In order to deal with the short-term disconnection between the DU and UE and the corresponding packet loss, we monitor the connection state of the UEs, and we generate traffic over them only when they are connected.

On the other hand, horizontal scaling follows a more complex strategy as mentioned in Section 5.4.2 since it requires deploying multiple CU-UP instances with different CU-UP IDs, which involves configuration updates and adjusting the node placement according to the

xApp's scheduling decision. For seamless and efficient horizontal CU-UP scaling, the system should support the CU-UP Change procedure, which enables UEs to handover between the CU-UP instances without the need for deregistration. Although this mechanism is defined in 3GPP Release 15 [12], it is not yet implemented in the current OAI software stack. As a workaround to this limitation, for the scale-out scenarios, we first deploy the additional CU-UP instances and subsequently restart the corresponding UEs so that they connect to the new CU-UPs. For scale-in scenarios, we remove the surplus CU-UP instances and restart the UEs connected to them, forcing a reconnection with the remaining CU-UPs. The drawback of this approach is the temporary disruption of UE connectivity, caused by container restarts, as well as potential packet loss during the short disconnection period. The CU-UP Change procedure as defined in [12] prevents packet loss by forwarding data from source to target CU-UPs. Since this is not yet supported, our implementation deals with packet loss following the same procedure as vertical scaling by monitoring the UE connection state and only generating traffic only when UEs are connected.



(a) UE1 iPerf Bitrate vs Demand (b) UE2 iPerf Bitrate vs Demand (c) Sum iPerf Bitrate vs Demands

Figure 6.6: Proactive Vertical CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate.

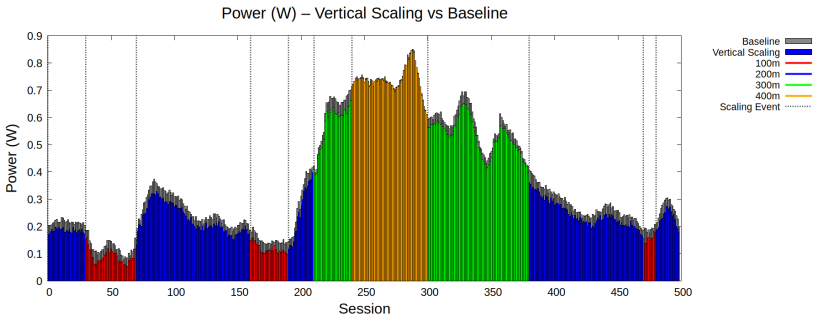


Figure 6.7: CU-UP Power Consumption Vertical Scaling vs Static Overprovisioning Across Sessions

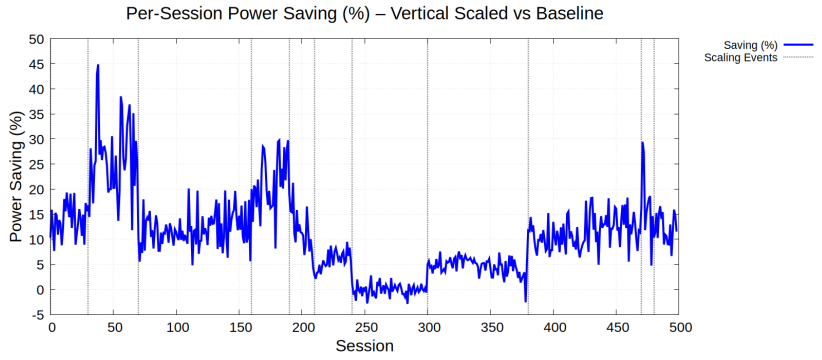


Figure 6.8: CU-UP Power Saving of Vertical Scaling against Static Overprovisioning

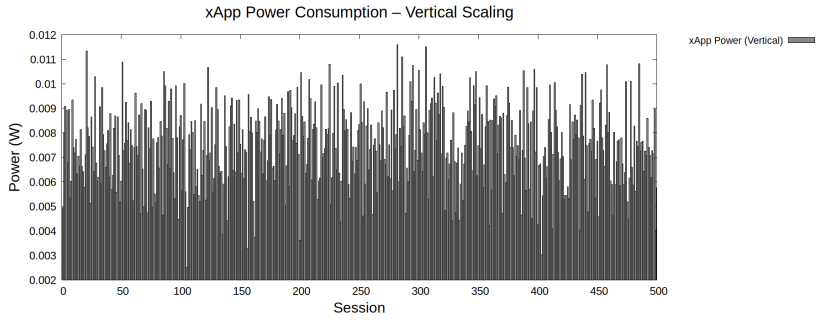


Figure 6.9: Power Consumption of Vertical Scaling xApp

Table 6.4: Maximum Sustainable Throughput per CU-UP CPU Provisioning Profile

CU-UP Profile (CPU Limit)	**Max Sustainable Throughput**
100 millicores (100mi)	50 Mbps
200 millicores (200mi)	130 Mbps
300 millicores (300mi)	210 Mbps
400 millicores (400mi)	300 Mbps

6.5.4 Evaluation

In this subsection, we describe the methodology followed in order to evaluate the proposed scaling approaches. While the CU-UP profiling process was introduced in Section 6.3, the resulting CU-UP capacity thresholds are summarized in Table 6.4 here to facilitate readability. More specifically, we compare the performance of the vertically and horizontally scaled systems against a baseline system.

Based on our experimental results, our 5G O-RAN system was able to achieve up to 130 Mbps stable Downlink data transmission rate per UE, while the uplink was limited to 60Mbps. Consequently, all the following evaluation experiments were conducted using Downlink traf-

fic so as to fully stress the CU-UP component.

In our setup, as described in Section 6.5.1, two UEs are connected, as a result, the maximum downlink throughput that the system can support is limited to 260 Mbps. Based on the analysis in Section 6.3 and as summarized in Table 6.4, the 400 mi profile is selected as the baseline configuration, since this resource level can sustain up to 300 Mbps, which is sufficient for a single CU-UP component to satisfy all traffic demands throughout the evaluation phase. The control interval used in Algorithm 4 and 5 was defined to be equal to the duration of 5 sessions. For all the energy measurements between a time period among t_1 and t_2, the total energy consumption is computed by integrating the instantaneous power consumption $p(t)$ over the evaluation interval, i.e., $E = \int_{t_1}^{t_2} p(t)\, dt$. Following a top-down approach, we evaluate the system's performance across the full execution, sub-peak periods, and specific traffic regimes.

Vertical Scaling

Figure 6.6 visualizes the achieved per UE DL throughput and the total aggregate throughput of the CU-UP Vertical Scaling system that utilizes the xApp presented in Algorithm 4 and the forecasting mechanism based on LGBM as presented in Section 6.4.1. Initially, the system operates with the 200mi profile for the first 35 sessions. This configuration is capable of satisfying the total served traffic, which exceeds the 50Mbps limit of the 100mi profile. In the following phase, the xApp predicts that the overall demand will decrease below 50Mbps and can be served by the 100mi profile. Once the first session appears when the 100mi profile can serve the traffic, the system triggers a scale-down operation, re-allocating resources while still satisfying the user demands and at the same time reducing the power consumption. During the next phases, the UE demands increase again. The xApp proactively scales up initially to the 200mi profile, subsequently to 300mi, and eventually to 400mi, each time selecting the profile capable of serving the demand that consumes the least power. When traffic decreases, the xApp scales-down the CU-UP instance, following the same principle of aligning resources with the traffic needs.

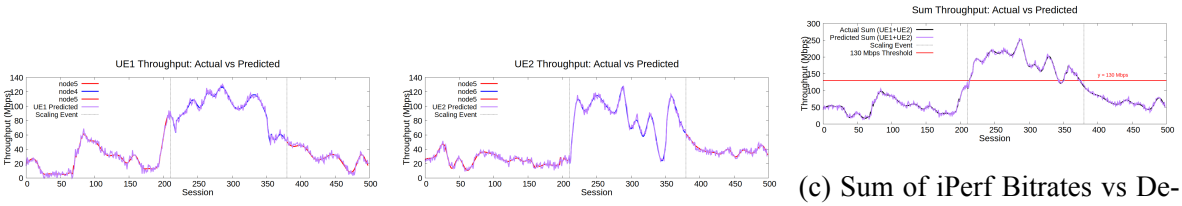
As seen in Figure 6.7, the system utilizing the vertical scaling algorithm described in Algorithm 4 manages to consume less power by employing the 100mi, 200mi, and 300mi profiles compared to the baseline system - baseline system relies solely on the 400mi profile throughout the entire traffic load experiment. This results in a total energy consumption

of 843.47 Joules for the vertical scaling system, compared to 920.32 Joules for the baseline system, achieving a 8.35% optimization during the entire execution. It is important to highlight that during the period that the system does not need the overprovisioned resources of baseline (excluding sessions 240-300), the vertically scaled system is limited to consume energy of 625.2J. As the baseline system consumes 702.7J, the achieved energy saving is approx. 11%. To better quantify the contribution of our proposed vertical scaling mechanism, we further categorize the experiment into four traffic regimes according to the required minimum profile to sustain the aggregated traffic: underutilized (100mi), nominal (200mi), high (300mi), and saturated (400mi) operation. As seen in Figure 6.6c, the Underutilized regime is observed during sessions 30–70, 160–190, and 470–480, where the aggregated traffic can be fully supported by the lowest provisioning profile of 100mi. The Nominal regime corresponds to sessions 0–30, 70–160, 190–210, 380–470, and 480–500. The High-load regime appears during sessions 210–240 and 300–380, while the saturated one appears between sessions 240–300. The energy aggregation per regime is summarized in Table 6.7. When traffic can be served by the lowest profile (Underutilized), the vertically scaled system achieves 23.6% energy savings compared to the static baseline. Under the nominal regime, savings remain substantial at 13.1%, while for the high-load regime the gain decreases to 6.5%. During saturated operation, energy consumption converges to the baseline system, resulting in zero difference, as expected.

The power consumption of the Vertical Scaling xApp can be seen in Figure 6.9. As illustrated, it requires energy consumption of 15.72J across all sessions. The xApp operation adds a small overhead to the overall system energy consumption, resulting in a 6.65% optimization during the entire dataset execution. During the period that the system operates without the overprovisioned resources, the overall energy consumption is 13.6J, resulting in 9.1% energy consumption savings.

Horizontal Scaling

Figure 6.10 presents the achieved per UE DL throughput and the total aggregate throughput from the Latency-Aware CU-UP Horizontal scaling system. This approach utilizes the xApp presented in Algorithm 5. In this approach, we use the 200mi profile as the CU-UP instances reference, since it can handle up to $\approx 130Mbps$, giving a balanced point between capacity and power consumption per data demands, while also being the middle point of the



(a) UE1 iPerf Bitrate vs Demand (b) UE2 iPerf Bitrate vs Demand (c) Sum of iPerf Bitrates vs Demands

Figure 6.10: Proactive CU-UP Scaling: iPerf Bitrates vs Demands for UE1, UE2, and Aggregate.

four tested profiles (100mi–400mi).

At the beginning of the experiment and between sessions 0-210, the system operates with a single CU-UP instance serving both DUs at Nodes 7,8. The CU-UP is placed on Node 5, as this results to the minimum CU-UP to DU RTT of 5ms. During this phase, the LGBM model accurately predicts that the aggregated demand and the resulting power consumption remain within the capacity limits of a single 200mi profile. After session 210, the xApp forecasts an increase in the total demand and the corresponding increase of the power consumption, exceeding the limit of a single 200mi CU-UP profile. Therefore, the scale-out mechanism is triggered to deploy a second CU-UP instance. With two active CU-UPs, the placement mechanism assigns one instance to Node 4 (serving DU Node 7) and the other to Node 6 (serving DU Node 8), achieving an RTT latency of 1.5 ms per CU-UP to DU link. A notable observation occurs around the 355 session, when the traffic falls below the 130 Mbps scale-in threshold. In this case, the xApp, predicts that the throughput will remain below the threshold only for a short period, and the system avoids an unnecessary scale-in operation and the corresponding ping-pong effect. But during the session 380, the xAPP predicts that the total traffic and the resulting power consumption will be below the scale-in threshold, and triggers the scale-in mechanism, resulting in a single CU-UP instance. This instance will be placed again at Node 5, since this placement results to the minimum CU-UP to DU latency of 5ms for both DUs. The latency-aware horizontally scaled system resulted in 877.36J of energy consumption during the entire experiment, which is 4.8% lower than the baseline system as seen in Figure 6.12 but 3.9% higher than the vertical one. If we also include in the discussion the energy consumption of the horizontal scaling xApp (Figure 6.13), which has been measured at 16.2J, the power saving against the baseline system results to 3% energy savings. In contrast, it spends 3.89% more energy compared to the vertically scaled system.

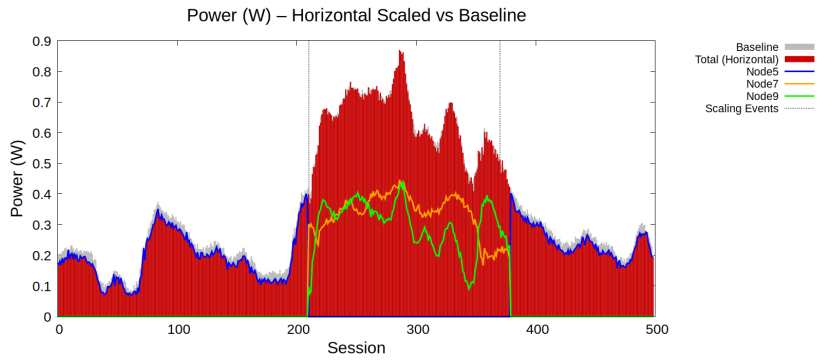


Figure 6.11: CU-UP Power Consumption Horizontal Scaling vs Static Overprovisioning Across Sessions

If we exclude the peak area where the system should operate with higher resources (sessions 240-300), the horizontal scaled system consumed energy of 648.56J over the 702.7J, succeeding in 7.7% energy saving. The power consumption of the xApp during the overprovisioned phases spent 14.1J, thus achieving an overall energy saving of 5.7%.

Following the same methodology as in Section 6.5.4, we split the experiment into the same four regimes according to the aggregated traffic. As shown in Table 6.7, in the underutilized regime, the horizontal scaling strategy achieves 16.61% energy savings compared to the static baseline. During nominal load conditions, savings remain noticeable at 12.76%. However, in the high-load regime, no energy savings are observed and drops to -2.42%, reflecting the activation of additional CU-UP instances to sustain the demand. During saturated operation, when peak capacity is required, the energy consumption converges to that of the baseline configuration. The activation of the second CU-UP in both the high and saturated regimes results to negligible differences compared to the baseline system. Putting everything together, Tables 6.5 and 6.6 contain the aggregate performance results of the baseline, vertical and horizontal scaling systems.

Comparison

Energy Saving As described in the previous subsections 6.5.4 and 6.5.4, both proactive scaling strategies manage to reduce the energy consumption of a 5G O-RAN compared to the baseline. Based on the metrics shown in Tables 6.5 and 6.6, together with the visual analysis of the power saving per scaling strategy (illustrated in Figures 6.11 and 6.14), it is confirmed that the vertically scaled system is more energy-efficient than the horizontally scaled one.

Specifically, during the entire experiment execution, the vertical strategy achieved an

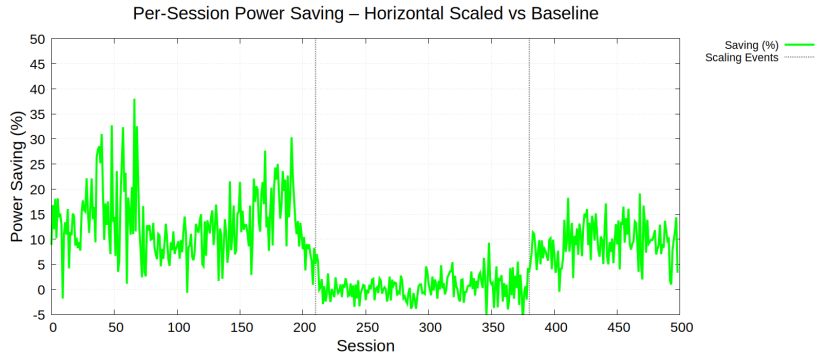


Figure 6.12: CU-UP Power Saving of Horizontal Scaling against Static Overprovisioning

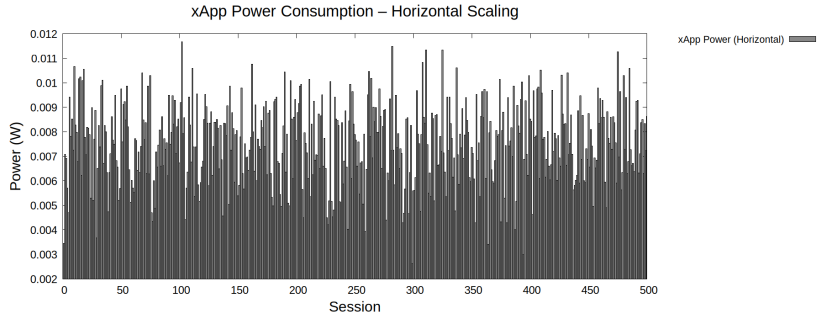


Figure 6.13: Power Consumption of Horizontal Scaling xApp

8.3% energy saving compared to the baseline, whereas the horizontal strategy achieved only a 4.7% reduction. This efficiency gap is further highlighted during the under-peak periods, as vertical scaling reached 11.0% savings compared to 7.7% for the horizontal approach. Although the full-execution energy savings of the two mechanisms are moderate (8.3% for vertical and 4.7% for horizontal scaling), this outcome is directly influenced by the traffic distribution of the evaluated dataset, which follows realistic UE behavior rather than artificially prolonged low-load conditions. As illustrated in Table 6.7, the most significant energy gains are observed during the underutilized and nominal regimes, reaching 23.6% and 13.1%, respectively. However, the evaluation is intentionally based on real-world traffic patterns instead of synthetic workloads that disproportionately emphasize off-peak intervals to enhance the real-world applicability of the results. Therefore, this approach limits the energy reduction. In practical O-RAN deployments with a larger proportion of underutilized or nominal operation, the energy savings would proportionally increase.

The developed xApps did not follow the energy-efficient RIC strategies proposed by Lima et al. [97], so the presented power consumption represents a baseline rather than an optimized result. Nonetheless, we still present the power overhead of the proposed xApps (see Figures

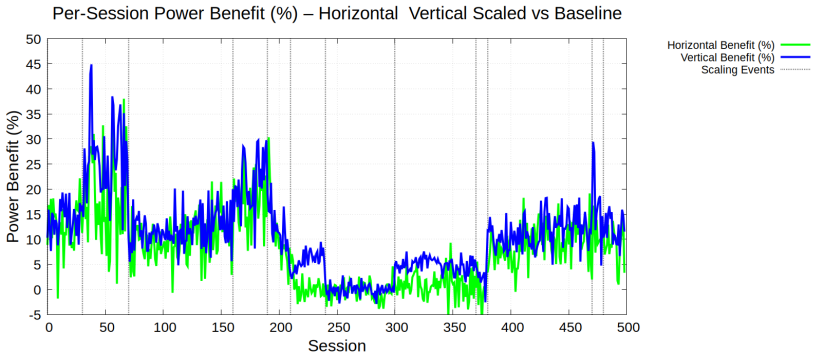


Figure 6.14: CU-UP Power Saving Comparison between Vertical and Horizontal Scaling System

6.9, 6.13) to provide a complete view of their contribution to overall 5G O-RAN energy consumption.

Table 6.5: Energy Consumption and Savings Over the Whole Execution

Scaling Strategy	**CU-UP (J)**	**xApp (J)**	**Total (J)**	**NFV (%)**	**System (%)**
Baseline	920.32	–	920.32	–	–
Vertical Scaling	843.47	15.72	859.19	8.3	6.65
Horizontal Scaling	876.44	16.20	892.64	4.7	3.0

Table 6.6: Energy Consumption and Savings During the Non-Overprovisioned Period

Scaling Strategy	**CU-UP (J)**	**xApp (J)**	**Total (J)**	**NFV (%)**	**System (%)**
Baseline	702.70	–	702.70	–	–
Vertical Scaling	625.20	13.6	638.8	11.0	9.0
Horizontal Scaling	648.56	14.1	662.6	7.7	5.7

Latency Figure 6.15 presents the mean RTT measurements in ms per session only for UE2, under the three tested environments: vertical and horizontal scaling, against the baseline system. It is important to note that UE1 and UE2 exhibited equivalent mean RTT behavior, and since the performance of UE1 is identical to UE2, only one UE RTT measurements are illustrated. Regarding the latency of the vertically-scaled and baseline scenarios, the RTT remains consistently stable with a mean value around 18 ms. This is happening since we empirically selected that Node 5 is used for deploying the single CU-UP instance. This was selected

Table 6.7: Energy Consumption per Traffic Regime: Vertical vs. Horizontal Scaling

Regime	Duration	Overprov (J)	Vertical Scaling			Horizontal Scaling		
			Energy	Saved	Saved (%)	Energy	Saved	Saved (%)
Underutilized	80	56.32	43.05	13.28	23.6	46.97	9.35	16.61
Nominal	250	327.71	284.64	43.07	13.1	285.92	41.79	12.76
High	110	314.98	294.37	20.61	6.5	322.60	-7.62	-2.42
Saturated	60	221.32	221.42	-0.09	-0.04	221.87	-0.55	-0.25
Total	**500**	**920.33**	**843.47**	**76.86**	**8.35**	**877.36**	**42.98**	**4.7**

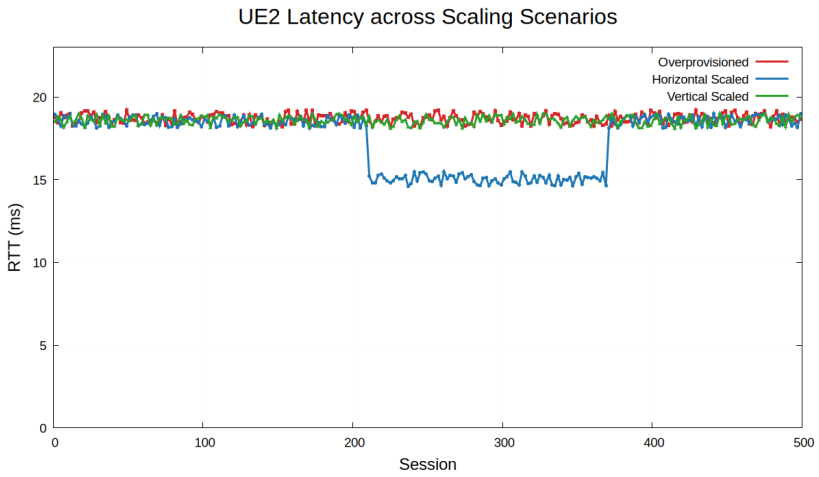


Figure 6.15: UE's Mean RTT under Different Scaling Configurations

as this node offers the minimum latency path over the midhaul link. During the first phase (sessions 0 - 210), no significant difference is observed between the two scenarios, since in both cases a single CU-UP instance is deployed, either vertically scaled or baseline. On the contrary, in the latency-aware horizontal scaling scenario, between sessions 210 and 380 the system scales out and operates with two CU-UP instances, placed at different nodes (Node 4 & Node 6). This placement reduces the CU-UP to DU RTT from 5 ms to 1.5 ms, thus leading to a small but important improvement, with the mean RTT being approximately 3.5 ms lower compared to the vertical and overprovisioned configurations as depicted in Figure 6.15.

To better understand the benefit of minimizing the latency of a UE and the corresponding 5G network slice, it is important to note that the packet delay budget defined by 3GPP as defined in [33] represents a strict upper bound rather than a performance target of QoS within a specific 5G network slice. This becomes particularly important for cases such as mixed-traffic slicing, which accommodate both eMBB users and real-time analytics flows. For example,

we can set a maximum RTT threshold of 20 ms for the aforementioned mixed-traffic slice. Although in the case of the vertical-scaling and overprovisioned network scenarios a single CU-UP being placed at Node 5 satisfies the baseline RTT constraint ($\approx 18ms$), it provides a small margin for end-to-end latency fluctuations caused by transient bursts, transport jitter and others. In contrast, in the case of the latency-aware horizontal scaled system, by scaling out and deploying CU-UPs at Nodes 4 and 6, the CU-UP to DU latency is reduced to $\approx 1.5ms$ per DU, resulting in a total end-to-end RTT of 15ms. Although both configurations satisfy the Service Level Agreement (SLA), the latency-aware horizontally scaled system leaves a margin of approximately 5 ms below the 20 ms limit, compared to only about 2 ms for the vertically scaled and overprovisioned configurations. This increases headroom and significantly enhances the robustness of low-latency flows against the frequent network dynamics of a 5G network, justifying the increase in energy overhead.

6.6 Discussion

As described in Section 6.5.4, the vertical scaling mechanism in an O-RAN environment enables more energy-efficient deployments, whereas the latency-aware horizontal scaling mechanism prioritizes latency optimization and achieves lower energy efficiency compared to vertical scaling, while still outperforming traditional overprovisioned baseline systems.

A key difference between the two scaling mechanisms is evident when we consider service continuity. In the experiments described in Section 6.5, traffic generation is initiated only after the scaling operation is complete and the UEs are re-attached to the network for both suggested approaches. However, in real-world deployments, vertical scaling is inherently more vulnerable to service disruption, as it typically requires restarting the CU-UP component, resulting in a short network downtime. In contrast, the horizontal scaling approach, when combined with the CU-UP Change procedure, enables UEs to be handed over to newly deployed CU-UP instances after they become operational, thereby preserving service continuity.

The vertical and horizontal scaling strategies have complementary characteristics. A hybrid CU-UP scaling strategy can optimize simultaneously the energy efficiency and the latency of the 5G O-RAN network. In a hybrid approach, vertical scaling could be used as the primary control mechanism during traffic variations, enabling energy-efficient resource adaptation within a single CU-UP instance. Vertical scaling will be applied until the system

approaches its resource limit threshold or when further resource increase becomes energetically inefficient. Beyond this point, the system transitions to horizontal scaling, and starts triggering the deployment of additional CU-UP instances. Once new instances are deployed, following latency-optimal placements, vertical scaling could be applied to each new CU-UP instance, allowing the system to converge toward an energy-optimal operating point.

Although the experimental evaluation was conducted with two DUs and three candidate CU-UP nodes, the proposed scaling mechanisms are designed for larger deployments. Moreover, as mentioned in Section 6.4.3, in larger deployments, the forecasting overhead can be further reduced by training the ML model directly at the DU level rather than aggregating per-UE predictions, thus reducing the number of forecasting calls from the total UEs to the number of DUs. In larger-scale systems, the number of connected UEs increases, the aggregate traffic exhibits a wider dynamic range between off-peak and peak periods, providing the scaling algorithms with greater optimization potential and consequently larger energy savings compared to static overprovisioning. Regarding the vertical scaling approach, our algorithm aggregates the total predicted demand of the number of DUs, so its operation remains unchanged as the topology grows. For horizontal scaling, the heuristic complexity is O(N·M), which scales efficiently with increasing numbers of DUs and candidate nodes, while larger topologies provide more degrees of freedom for latency-optimal CU-UP placement.

Our current implementation requires the cooperation of multiple ML modules, which increase system complexity and operational overhead. By introducing a Foundation Model (FM), our proposed system's complexity can be reduced, since FMs are able to unify multiple tasks into one single model. Given historical KPIs, network topology information, and energy constraints, an FM can directly infer the appropriate scaling and placement action, eliminating the need for multiple ML pipelines and handcrafted heuristics.

6.7 Conclusion

In this work, we have designed and implemented a real-world 5G O-RAN network architecture featuring energy-efficient proactive scaling strategies for the CU-UP component and evaluated them on a testbed environment under dynamic network conditions.

Initially, an experimentally validated CU-UP power model was developed using power monitored measurements, demonstrating that allocating the minimum sufficient CPU resources

consistently minimizes power consumption under varying traffic loads. Two proactive scaling strategies driven by ML-based traffic demand forecasting were designed and implemented as Near-RT RIC xApps: namely the vertical CU-UP scaling and the latency-aware horizontal CU-UP scaling.

A critical observation from our study is that energy savings are inherently dependent on the underlying traffic patterns. Specifically, power reduction cannot be achieved when the system operates under peak demand, as the maximum computational resources must be allocated to sustain the required throughput and prevent service degradation. Based on our experimental results we defend that the vertical scaling strategy maximizes the energy efficiency since the system achieves up to 11% energy savings during sub-peak periods, up to 23% during underutilized regimes and 8.3% over the full execution compared to the static baseline system. On the contrary, the horizontal scaling introduces a complementary trade-off, delivering lower energy savings (up to 7.7% in non-peak periods, 16.61% during the underutilized regime and 4.7% for the whole execution) while significantly reducing the 5G-RAN latency by performing latency aware placement of the scaled CU-UP instances, thereby increasing E2E latency headroom and improving the robustness for the latency-sensitive services. Additionaly, it is important to highlight that the measured xApp overheads remain low relative to CU-UP energy consumption, confirming that proactive RIC-based control is feasible and beneficial in operational O-RAN networks.

Overall, this work demonstrates that traffic-aware CU-UP scaling is a practical and effective lever for improving the sustainability and energy efficiency of O-RAN deployments, highlighting how O-RAN's intelligent automation based on the RIC-driven programmability enables energy-aware control and optimization in 5G networks. Future work will focus on the design of a hybrid CU-UP scaling strategy utilizing Foundation Models that jointly combine the elasticity of the vertical and the horizontal strategies. Despite addressing the energy efficiency of the CU-UP, the power consumption of the DU and RU components, which represent the dominant share of the total RAN energy footprint, remains unaddressed and is the focus of Chapter 7.

Chapter 7

KPI-Driven Power Consumption Prediction and Anomaly Detection in O-RAN Networks

7.1 Introduction

Chapter 6 demonstrated that proactive CU-UP scaling can reduce the energy consumption of the 5G RAN UP. CU-UP represents only one component of the O-RAN energy footprint, while the DU and RU, which are responsible for baseband processing and RF transmission, respectively, account for the dominant share of the total RAN power consumption [10]. Understanding and monitoring the power behavior of these components is therefore a prerequisite for any energy-aware optimization targeting the DU and RU.

The operational state of each O-RAN component can be reflected through the Key Performance Indicators (KPIs) and the corresponding Key Performance Measurements (KPMs), which implement protocol-level telemetry and originate from different layers of the stack. More specifically, the RU exposes metrics that are related to the physical layer and the RF transmission, such as Physical Resource Block (PRB) utilization and signal quality, while the O-DU and O-CU expose higher-layer telemetry such as HARQ acknowledgment, scheduling statistics, user-plane throughput and others. These KPIs are used for resource management and network optimization through decisions made by the RAN Intelligent Controllers (RICs). The Near-Real-Time (Near-RT) and Non-Real-Time (Non-RT) RICs utilize KPIs for short- and long-term RAN optimizations, respectively. In the evolving landscape of beyond-5G and

6G networks, most works primarily focus on using KPIs to enhance network intelligence and orchestration. For instance, KPIs have been widely used for traffic forecasting, traffic steering, and network slicing optimization [5], [98], and [99]. However, an important and largely unexplored question remains as depicted in **RQ5**: **To what extent can network KPIs predict the power consumption of O-RAN components? And how can they be used for anomaly detection and calibration?**

Answering this question is crucial for establishing sustainable O-RAN deployments. In the current O-RAN deployments, power monitoring is decoupled from the control plane and depends on software-based solutions or external hardware sensors, which typically suffer from calibration drift and anomalies. Additionally, the impact of the O-RAN network measurements to the power consumption of the O-RAN components is underexplored due to vendor-specific hardware diversity, or is limited to works that rely on simulation data or energy models that in practice prove to be far from reality.

In this chapter, we perform an extensive study on the predictability of power consumption using network KPIs within a B5G/6G experimental research infrastructure. We focus our analysis on the O-RU and O-DU, as these components represent the highest-energy-intensity components of the RAN stack. In this context, we propose a framework that aims to deliver an AI-driven inference engine that performs real-time calibration and anomaly detection of the power consumption measurements for the O-RAN deployment, relying solely on the reported KPIs and intentionally excluding historical power values (W) to prevent measurement 'poisoning' with sensor inaccuracies.

Our methodology follows three different stages. First, we conduct a data collection campaign on the experimental infrastructure, using multiple scenarios that generate dynamic traffic patterns up to 900 Mbps, to capture the relationship between reported KPIs and DU/RU power consumption. This collected dataset is subsequently processed as a time series, appropriately prepared and filtered. The resulting metrics are used to train node-specific AI-driven power predictors, using a hybrid XGBoost model for the O-RU and a hybrid XGBoost/CatBoost model for the O-DU. As a second stage, calibration and anomaly-detection headers are designed and deployed to analyze the residuals between predicted power consumption and actual hardware meter readings. By evaluating the absolute percentage error against node-specific thresholds, the system is able to identify measurement drift or sensor inaccuracies. When an anomaly is detected, the framework executes autonomous calibration by

replacing the anomaly hardware metric with the predicted value. Lastly, a centralized monitoring service is integrated into the architecture to offer real-time observability via a web-based dashboard. During the evaluation phase, the proposed hybrid models achieve low prediction errors, with Mean Absolute Error (MAE) values around 0.97 W for the O-DU and about 0.58 W for the O-RU, demonstrating accurate power estimation from network KPIs. The anomaly detection mechanism achieves a 100% detection rate for manually injected power measurement anomalies, followed by a calibration header that removes the detected anomalies and replaces them with the predicted power consumption values from the ML models.

Putting everything together, the main contributions of this chapter are:

- To analyze the correlation between KPIs and the power consumption of O-RU and O-DU.
- To design and develop ML models for predicting node-level power consumption using O-RAN KPIs.
- To design and develop an AI-driven framework for validating and calibrating energy measurements in O-RAN systems.
- To experimentally validate the proposed system on a real O-RAN testbed.

7.2 Related Work

Recent research studies have explored the usage of KPIs to enable intelligent management of B5G & 6G networks. Several prior works [5], [98], and [99] focus on the reported O-RAN KPIs, which are analyzed and subsequently used for traffic forecasting and scaling of O-RAN components, traffic steering, and network slicing optimization. Authors in [100] utilize KPIs to improve network mobility management through predictive handover mechanisms. Work in [101] presents the usage of platform based KPIs and O-RAN KPIs to predict different kind of anomalies such as CPU contention, Radio interference, network contention and a mix of anomalies.

Focusing more on power consumption, different papers aim to create power consumption explainable models of the O-RAN networks components, based on their behavior. For instance, authors in [102] create a nonlinear power model for the O-DU that predicts power consumption based on traffic load and CPU utilization. The work is limited to theoretical

formulation and simulation-based evaluation, without leveraging real-world KPIs or operational measurements. A more experimental approach is followed in [103], where the authors experimentally evaluate the power consumption of virtualized monolithic gNBs and derive statistical power consumption models based on measurements that reflect O-RAN compatible KPIs such as SNR, MCS, Traffic load and others.

Valid and calibrated power consumption metrics are becoming increasingly important in the O-RAN architecture, witnessed by the fact that several research papers focus on minimizing the power consumption of the O-RAN components. For instance, authors in [88] propose an energy-efficient placement of O-RAN components and the corresponding user association based on power consumption metrics. The calibrated and anomaly-free power consumption measurements are also important for validating energy efficiency optimization techniques, such as the evaluation of energy savings from Advanced Sleep Modes of the RU presented in paper [104], as well as dynamic resource allocation policies for O-RAN components evaluated through real-world energy measurements in [9].

Despite the aforementioned progress in KPI-driven network management, power consumption modeling, and energy-efficient optimization techniques for O-RAN networks, the reliability of power telemetry remains largely unexplored. Inaccurate or anomalous power measurements can significantly affect energy-aware optimization and evaluation results. To address this limitation, our work proposes AI-Driven Approach Using Network KPIs to predict the power consumption of O-RAN components to calibrate and detect anomalies in the reported power measurements.

7.3 System Architecture

The proposed system architecture integrates a data-driven monitoring and inference framework for real-time energy prediction, calibration, and anomaly detection into an O-RAN 5G network, as illustrated in Figure 7.1.

The system architecture is based on a typical 5G private network, which is deployed at the VTT 5GTN O-RAN testbed, following fully disaggregated O-RAN deployment decomposed to a Central Unit (CU), Distributed Unit (DU), and Radio Unit (RU). Two separate nodes were used to host the CU and DU components, based on the Accelleran O-RAN software stack, deployed as containerized services on commercial off-the-shelf servers. The CU and DU

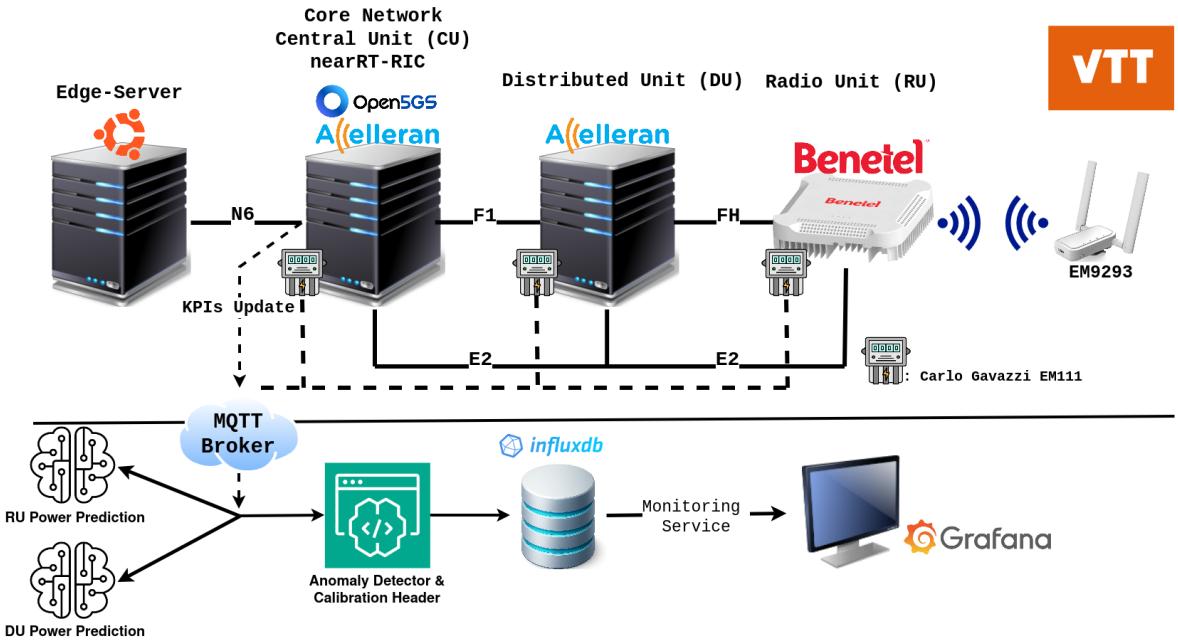


Figure 7.1: System Architecture and O-RAN Testbed Deployment

communicate via the F1 interface, enabling separation of higher- and lower-layer baseband processing. The radio interface is provided via a real-world radio unit device, namely the Benetel RAN650 O-RU, following the O-RAN 7.2x functional split and connects to the DU through the fronthaul (FH) interface. Each of the disaggregated O-RAN components operates as an E2 node connected to the near-RT RIC, hosted at the same server Utilizing the E2SM-KPM model, the O-RAN components periodically report multiple KPMs to the near-RT RIC. For the UE, a commercial laptop equipped with a Sierra Wireless EM9293 5G modem is used to connect to the O-RAN network.

For the 5G Core Network we used the Open5GS software suite, where the VNFs are deployed on the same host with the CU and nearRT-RIC components as containerized services, including the Network Repository Function (NRF), Access and Mobility Management Function (AMF), Session Management Function (SMF), User Plane Function (UPF), Authentication Server Function (AUSF), Unified Data Management (UDM), Unified Data Repository (UDR), Policy Control Function (PCF), and the Network Slice Selection Function (NSSF). These functions provide the necessary control-plane and user-plane services required for end-to-end 5G connectivity. One more node has been used to serve as the Data Network gateway, connecting the 5G CN & O-RAN network to the global Internet. The same node operates as the testing traffic generator to the 5G network.

A dedicated xApp is deployed at the Near-RT RIC, and is responsible for exporting KPI

reports from the DU and RU O-RAN components through the E2SM-KPM service model. In addition, real-time power consumption measurements of the O-RAN components are obtained from Carlo Gavazzi EM111 energy analyzer devices installed on the physical nodes hosting the O-RAN components, providing power monitoring at a 1 Hz sampling rate. The KPI reports and energy measurements are streamed via an MQTT-based telemetry layer, enabling the transfer of monitoring data from the 5G O-RAN infrastructure.

This chapter presents experiments on a single 5G O-RAN deployment. Nevertheless, the proposed AI-based framework for power prediction, calibration, and anomaly detection is designed following a multi-site architecture, where the MQTT telemetry layer connects the 5G O-RAN site (located at VTT testbed) with the remote AI components (at our premises). To support data analysis, a dedicated InfluxDB time-series database is deployed at our AI site to store collected KPIs and power-consumption measurements. The first component of the AI workflow consists of two inference services that host pre-trained machine learning models (presented in Section 7.4) for power prediction. The calibration and anomaly detection headers compute the residual between the measured and predicted power for each O-RAN component. If the residual error exceeds a node-specific threshold (e.g., 5%), the system flags an anomaly, indicating possible sensor drift, hardware failure, or unexpected load. When this occurs, the framework automatically recalibrates by replacing the faulty hardware metric with the predicted value to maintain reliability. This functionality is illustrated in Figure 7.2. The inference components are containerized to ensure reproducibility and facilitate an easy-to-deploy architecture. In addition, the system includes a monitoring service implemented with Grafana that provides real-time visualization of network KPIs, power measurements, and prediction results.

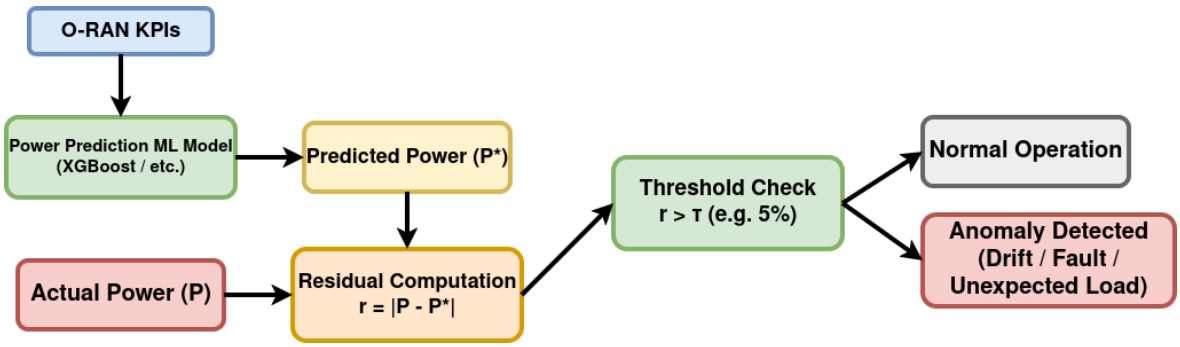


Figure 7.2: Residual-Based Power Anomaly Detection Architecture

7.4 Methodology & AI Model Design

As already mentioned, our proposed system architecture, constitutes of two independent ML models trained on the collected data, performing predictions on the power consumption for each node that hosts the O-DU and O-RU.

7.4.1 Dataset

The dataset used to train the proposed ML models exhibits a recurring daily pattern: the Downlink (DL) bitrate gradually increases from 0 to 900 Mbps (the system's maximum supported traffic) and then decreases in the same manner. Each transmission lasts 10 seconds at a constant rate, and this pattern is repeated over 7 consecutive days. This dataset is ideal for our ML model training, since it provides sufficient duration to effectively capture the power consumption behavior of the nodes and the 5G components, and it has representative system behavior. It captures the full operating range of the system, and the power consumption measurements are correlated with traffic levels, enabling the models to learn not only the temporal dynamics of traffic evolution but also the nodes' energy behavior under varying load conditions. Additionally, the KPIs from the AcceleRAN platform allow training beyond DL bitrate, incorporating metrics such as ACKs/NACKs and buffer sizes. These features provide deeper insights into the behavior of the 5G radio components (DU, RU), as confirmed by the correlation analysis that follows.

7.4.2 Correlation Analysis

Table 7.1: Comparison of DU and RU KPIs Correlation with Power Consumption (W)

KPI	**DU Correlation**	**RU Correlation**
DlDataHARQAcks	0.952555	0.990073
UlDataHARQAcks	0.937408	0.978520
DlDataRate	0.866069	0.760212
DlPrbsUsed	0.860628	0.752793
DlDataHARQNacks	0.848326	0.859766
AvgPucchSnr	0.806731	0.849290

DU

A DU component in a 5G-RAN network is responsible for the digital baseband processing, integrating the following functionalities: scheduling, HARQ handling, channel coding/decoding, and resource allocation. Its power consumption is related to the amount of traffic and control signaling it processes, without being responsible for the RF transmission. The DlDataHARQAcks reflects the successful DL HARQ transmissions, resulting to the volumes of the user plane data that are processed from the DU. More acknowledgments lead to more encoding rounds, more packets to schedule, and more rounds for resource allocation, ultimately increasing the baseband processing load. The UlDataHARQAcks correlate with the power values (W) due to the UpLink (UL) control signaling from the UEs. Of course, DLDataRate reflects the measured throughput of the data volume the DU must process. Higher DL rate results in higher symbol mapping, coding, and multiplexing operations, which in turn creates increased computational intensity. The DlPrbsUsed metric represents the number of Physical Resource Blocks (PRBs) scheduled in the DL, which directly reflects spectrum occupancy. A higher number of PRBs leads to increased baseband operations and more per-user resource mapping, thereby intensifying the baseband activity and raising DU power consumption. Similarly, DlDataHARQNAcks captures the number of DL retransmissions triggered by NACKs. These retransmissions introduce additional buffer usage, scheduling complexity, and channel-coding overhead, all of which increase the DU's computational load and energy cost. Finally, AvgPucchSnr reflects the quality of the physical UL control channel (PUCCH). As SNR improves, ACK/NACK control messages are more reliably decoded at the DU, leading to increased HARQ-related processing activity. Although its correlation is weaker than that of pure traffic KPIs, UL control still significantly contributes to DU workload, as it directly drives HARQ feedback cycles. Given the operational mechanics described, it is evident why these specific KPIs exhibit a high correlation (> 0.8) with power consumption and the reason why these metrics were selected as the optimal features for training the DU power predictor.

RU

The DlDataHARQAcks metric shows the strongest correlation with the RU power, since every successful DL HARQ delivery requires RF transmission. Subsequently, more packets reflect to larger durations that the RU Power Amplifier (PA) and RF chains are active.

Although the UlDataHARQAcks metric looks fuzzy since no UL transfer was running, this high correlation reflects the UL HARQ feedback (control) from the UE. The DlDataHARQ-NAcks metric indicates retransmissions. When errors happen, the RU must re-transmit the same packets, which increase the active time of the PA and the RF chain, adding to the energy cost. The AvgPucchSnr reflects the quality of the UL control channel. When SNR is high, the RU can support a higher rate of ACK/NACK exchanges and the radio front–end stays engaged in decoding, hence more correlation with the actual consumed power. On the other hand, DlDataRate and DlPrbsUsed also impact the RU, since higher throughput and more PRBs indicate more continuous RF transmissions. Nevertheless, the correlation is lower because modulation/coding efficiency can hide some of the power variation. Based on the aforementioned metrics, although some KPIs have smaller correlation comparing to the DU case (> 0.7) we chose to include them in the feature set since they capture the essential physical layer and RF transmission dynamics that directly impact the energy footprint of the RU.

7.4.3 Machine Learning Models

It is important to mention that in this work, power Forecasting does not follow the typical time-series conventions as done in [105]. On the contrary our proposed AI models were trained exclusively on network KPIs, in order to avoid the propagation of poisoned data from faulty hardware sensing.

DU

Focusing on the DU, we trained individual models using only the raw KPIs that show correlation higher than 0.8 with the power consumption. This initial setup served as our baseline, allowing us to observe the behavior of popular regression models such as Random Forest, XGBoost, CatBoost and Autoencoders under purely KPI-driven input. The achieved results, with MAE values above 6 W, confirmed that the models can capture only a limited part of the consumption dynamics. Following that, we apply feature engineering by introducing a TrafficActive flag to distinguish idle from traffic phases, which slightly improved the performance, validating that phase-awareness adds temporal context. The next step is to include the differential features (ΔKPIs) as features to represent short-term variations in the KPIs and how these differences could affect power consumption. Combining ΔKPIs with TrafficAc-

tive further enhances generalization, demonstrating that temporal dynamics, together with traffic awareness, better describe actual DU behavior.

Since the DU uses a general-purpose CPU, its baseline power consumption varies with system load and background activity, leading to deviations between training and inference baselines. To address this, we introduce a baseline offset correction that dynamically tracks and adjusts to these variations, keeping the model aligned with the actual DU behavior. Based on that, we experimented with creating two different models, each of them explicitly trained to separate idle and active states. In these models, lightweight regressors (LR, XGBoost, RF) manage the stable idle phase, while CatBoost is employed during traffic periods. The inclusion of a long-term (1800 s) baseline further stabilizes the predictions, leading to our best-performing configuration, namely the Hybrid (Idle = XGBoost, Traffic = CatBoost) approach, which achieves MAE = 0.97 W and $R^2 = 0.946$.

RU

For the RU, we follow a similar progressive strategy, aiming to train models that can capture the temporal dynamics and context behind its power behavior. We evaluated several architectures, including Random Forest, CatBoost, and the XGBoost model. We started with models based on the basic KPIs, combined them, then applied feature engineering to add the TrafficActive indicator, and gradually expanded the feature space by introducing ΔKPIs and lag-based KPI metrics to better model short-term variations in RU power consumption. This analysis, summarized in Table 7.2, led us to select a hybrid XGBoost model (XGBoost is used for both idle and active) with KPIs + lag1 + ΔKPIs TrafficActive as input features, as it demonstrates the best overall performance of MAE = 0.581 W and R^2 = 0.885.

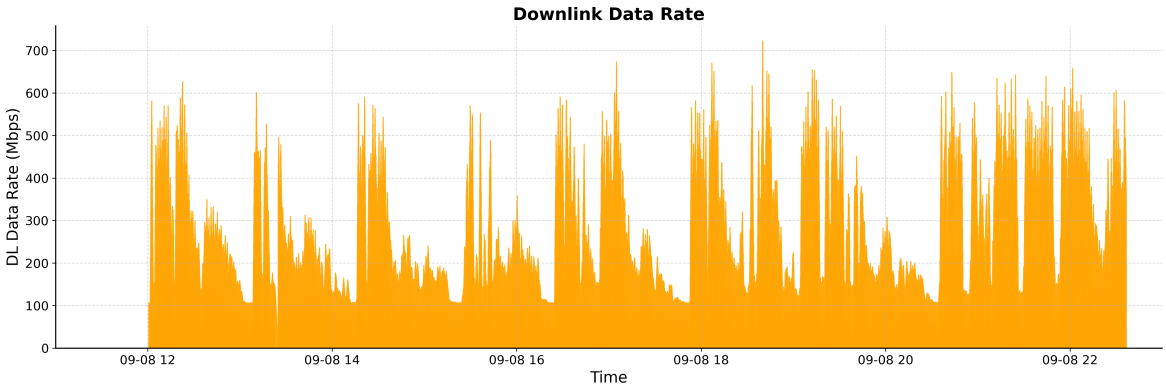


Figure 7.3: Generated DL DataRate during experiments

Table 7.2: Comparison of Top 5 Performance Models for DU and RU Components

Node	Rank	Model Architecture	Key Features	MAE (W)	R^2
	1	Hybrid (XGBoost/CatBoost)	KPIs + ΔKPIs + 1800s baseline	0.97	0.946
DU	2	Autoencoder	Basic KPIs + ΔKPIs + TrafficActive	1.96	0.929
	3	Hybrid (RF/CatBoost)	KPIs + ΔKPIs + 1800s baseline	1.37	0.902
	4	Autoencoder	Basic KPIs + ΔKPIs	2.19	0.869
	5	CatBoost	Basic KPIs + ΔKPIs + TrafficActive	1.92	0.852
	1	XGBoost (Idle, Traffic)	KPIs + lag1 + ΔKPI + TrafficActive	0.581	0.885
RU	2	CatBoost (Idle, Traffic)	KPIs + lag1 + ΔKPI + TrafficActive	0.783	0.870
	3	Random Forest (Idle, Traffic)	KPIs + lag1 + ΔKPI + TrafficActive	0.890	0.815
	4	Random Forest	Basic KPIs + TrafficActive	2.56	0.577
	5	CatBoost	KPIs + ΔKPIs + TrafficActive	1.625	0.563

7.5 Experimental Results & Validation

To validate our proposed framework, we conducted extensive experiments by generating UDP DL traffic streams using iperf based on the real-world dataset described in Section 2.7 over the VTT testbed. Figure 7.3 depicts the patterns in the replayed dataset.

To validate the performance of the RU and DU models, we conducted a 12-hour inference experiment (11:00–23:00 UTC) using the corresponding ML models. Figure 7.4 presents the predicted and measured RU power consumption over time. As shown in Figure 7.4, the forecasted power succeeds a strong correspondence to the actual Carlo Gavazzi EM111 measurements, achieving a MAE = 0.652 W, Mean Absolute Percentage Error (MAPE) = 1.02%, Root Mean Square Error (RMSE) = 1.095 W, and R^2 = 0.8254, indicating highly accurate alignment between the model output and real energy measurements, with an average infer-

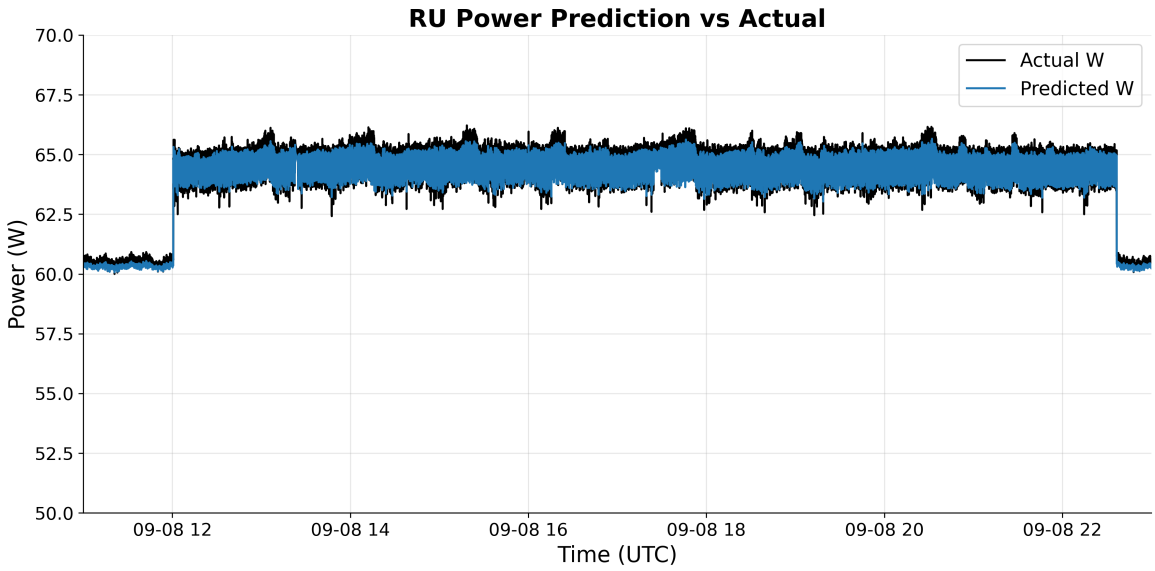


Figure 7.4: RU Power Forecast vs. Measured Power

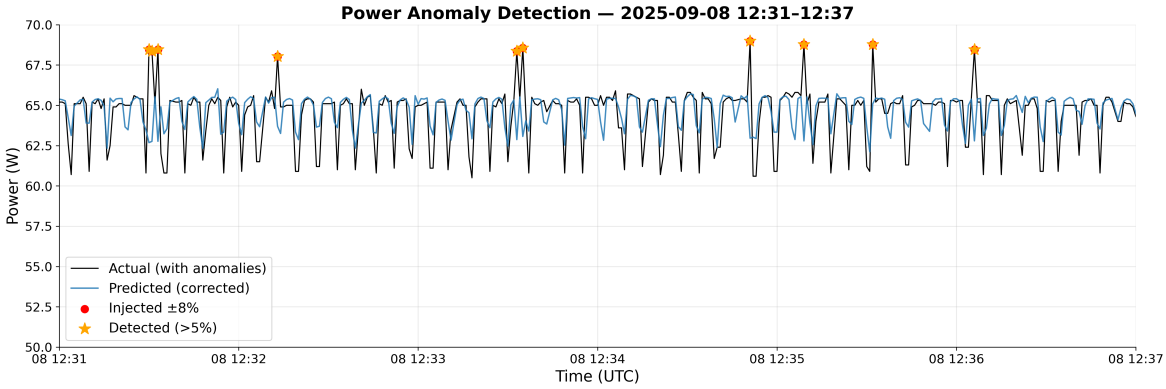


Figure 7.5: Detection of Injected +8% Power Anomalies on the RU Node

ence time of 1.13 ms per prediction. To evaluate the anomaly detection mechanism, 10 artificial anomalies were injected in the actual measurements for the time duration of 12:31 - 12:37 as shown in Figure 7.5, by increasing the measured value of W by 8% in 10 random timesteps. The model predicted values were compared with the perturbed measurements, and the residuals, defined as the absolute percentage error between the two, were analyzed. All injected anomalies produced residuals above the selected baseline of 5% threshold and were successfully detected, achieving a 100% detection rate.

Figure 7.6 shows the predicted power values compared with the actual power measurements of the DU component. As we can observe, during the idle phase, the predicted DU power appears much more stable compared to the measured one. This behavior occurs because the power variations observed in the measurements are not driven by network KPIs

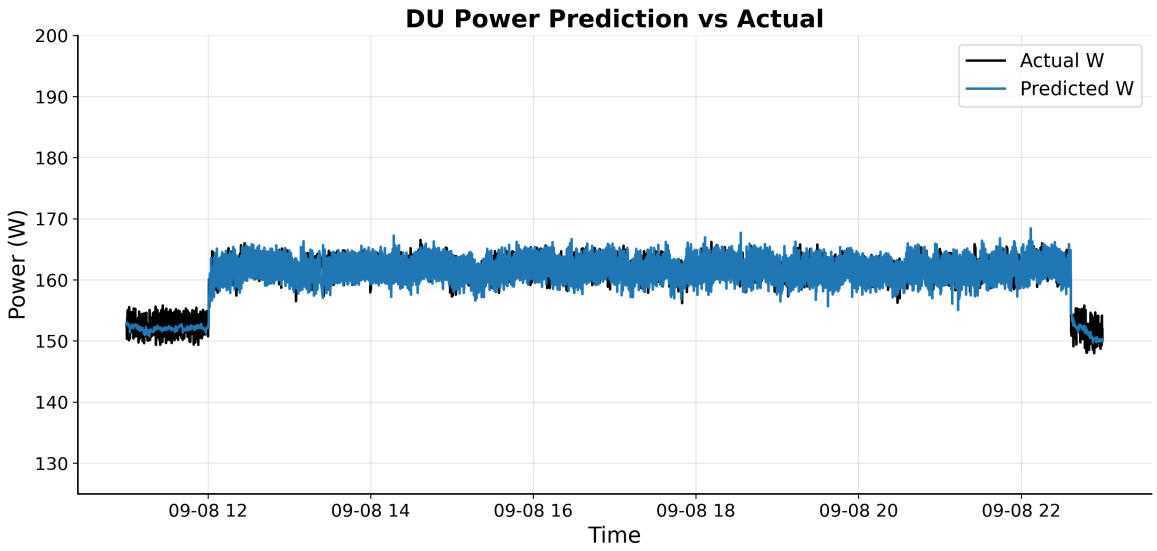


Figure 7.6: DU Power Forecast vs. Measured Power

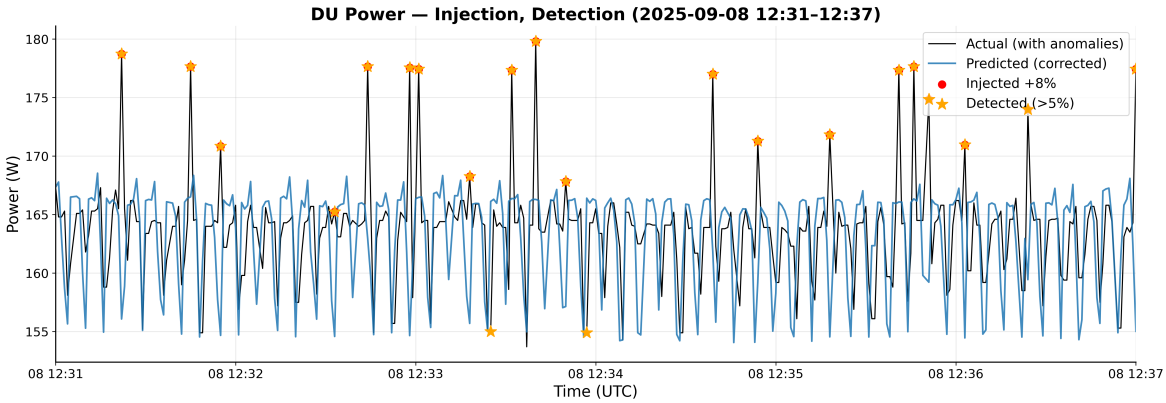


Figure 7.7: Detection of Injected +8% Power Anomalies on the DU Node

but mainly by background noise from the DU's CPU. Additionally, during the traffic periods, the model follows the actual power measurements closely but appears slightly more aggressive during the short idle phases (as shown in Figure 7.7). This behavior occurs because the idle intervals between consecutive traffic transmissions are extremely short, typically lasting around one second. During these brief spikes, the model consistently predicts lower power values than the measured ones, because the actual drop in power consumption during such a short idle period is largely arbitrary and dominated by the transient behavior of the DU's CPU rather than by the network state itself. During the validation phase, the model achieved an MAE = 1.954 W, MAPE = 1.22%, RMSE = 2.120 W, and $R^2 = 0.735$, with an average inference time of 1.2 ms per prediction.

Following the same methodology applied for the RU, the anomaly-detection mechanism

was evaluated by injecting 10 artificial anomalies into the actual power measurements over the same observation window. As shown in Figure 7.7, these anomalies were introduced by increasing the measured power W by 8% at 10 randomly selected timesteps. The model successfully identified all injected anomalies against the 5% baseline threshold and, notably, detected two additional genuine anomalies originating from the raw power measurements.

7.6 Conclusion

In this work, we explored the feasibility of predicting the power consumption of O-RAN components using only network-related KPIs. Experiments on a real 5G O-RAN testbed showed that KPI-driven ML models can accurately estimate the power consumption of both O-DU and O-RU nodes, achieving errors below 2W. Based on these predictors, we designed a lightweight framework that performs residual-based anomaly detection and automatically calibrates hardware power measurements. The evaluation confirmed reliable detection of injected anomalies and real-time inference.

For future work, we aim to extend the framework by designing a mechanism for the CU component, enabling a complete AI-based anomaly detection and calibration framework across all O-RAN components. Additionally, the ML-based power prediction mechanism will be deployed at the Near-RT RIC as an xApp, allowing closed-loop control and energy-aware RAN optimization.

Chapter 8

Conclusion

This thesis investigated the integration of 6G era's vision towards intelligent, autonomous, and energy-aware RAN operations by performing dynamic control and management of the 5G disaggregated RAN. Specifically, it addresses four 6G service categories and one global design direction defined by the ITU-R IMT-2030.

The UC service category is addressed by extending indoor coverage through heterogeneous RAT integration, while the IMC is explored via elastic RAN UP scaling to meet the high bandwidth demands of next-generation applications. The HRLLC is targeted through latency-aware placement of the scaled RAN UP components, reducing the end-to-end latency for latency-sensitive services. The IAC is followed by embedding ML-driven intelligence natively into the RAN control plane via O-RAN RIC controllers. Finally, the Sustainability design direction is enhanced through energy-aware orchestration and power-monitoring mechanisms that minimize the energy footprint of disaggregated RAN deployments.

8.1 Summary of Contributions

Thesis research began by solving the fundamental cellular problem of indoor coverage. Real-world 5G deployments operate exclusively on licensed spectrum, suffer from indoor coverage limitations, and undermine the 6G vision for the UC service category. **RQ1** asked how non-3GPP access technologies could be integrated directly into the disaggregated RAN stack to enable unified capacity extension and service differentiation at the access layer. As described in **Chapter 3**, the first contribution is addressing the aforementioned coverage problem by integrating a WiFi DU into the disaggregated 5G RAN stack. This RAN-level

convergence extended the system capacity without additional licensed spectrum. Alongside this architectural integration, a user-based slicing mechanism was built into the WiFi MAC layer via SDN and MAC Access Categories, enabling per-user QoS without modifying the standard 5G CP signaling. The experimental validation confirmed the viability and advantages of this approach, demonstrating improved system capacity and service differentiation in a real RF-isolated testbed environment.

Being able to deploy a 5G multi-RAT RAN system immediately raised concerns about static policies that cannot respond to the dynamic and heterogeneous nature of different wireless channels. So the **RQ2** was asking how AI-driven control mechanisms within the O-RAN architecture could enable proactive and intelligent traffic steering in a heterogeneous multi-RAT system. **Chapter 4** is answering this RQ by elevating the multi-RAT architecture into a fully O-RAN-compliant system that introduces intelligent and proactive traffic steering via xApp deployed on the Near-RT RIC responsible to take the traffic steering decisions. The xApp used an LSTM-based demand forecasting model to anticipate when the 5G DU will become saturated and offload traffic to the WiFi DU before congestion occurred. Based on the real-world experiments conducted, the results confirmed that this proactive approach can meet high-throughput demands and ensure fairness even under high-contention conditions. These research outcomes revealed that as multi-RAT utilization grows, the aggregated traffic volume processed by the CU/CU-UP increases substantially, making the CU-UP the primary bottleneck in the RAN user plane.

Based on the aforementioned, the fact that the CU-UP becomes the primary bottleneck in the RAN UP motivates the third contribution. In this context, the **RQ3** asks how cloud-native RAN UP functions can be autonomously scaled in an O-RAN deployment to maintain service performance under dynamic traffic conditions. **Chapter** 5 proposes a proactive horizontal scaling mechanism that reuses the LSTM forecasting model from the previous contribution to anticipate the future client demands and perform scaling out/in of the CU-UP component. This proactive horizontal scaling mechanism was compared with the baseline and the corresponding reactive horizontal scaling mechanism. Reactive scaling is fundamentally constrained because scaling decisions are made after congestion has already occurred, resulting in a service downgrade. On the contrary, proactive scaling eliminates this gap and consistently outperformed its reactive counterpart in terms of QoS continuity. However, every scaling decision carries a cost: each additional CU-UP instance provisioned to sustain

performance increases the deployment's energy footprint, and this tension between service quality and energy efficiency becomes the central question of the next contribution.

RQ4 asked how different RAN UP scaling strategies could be designed to jointly optimize performance, latency, and energy efficiency in disaggregated O-RAN deployments. **Chapter 6**represents the fourth contribution, therefore redesigning the scaling strategies aiming to jointly optimize performance, latency, and energy efficiency. To realize these strategies, an experimental source power model was first derived from real testbed measurements, capturing the relationship between the CU-UP CPU resources and the power consumption across different traffic conditions. Two different scaling strategies have been designed and deployed at the O-RAN near-RT RIC. The first implements a vertical scaling approach, dynamically adjusting the CPU resources of a single CU-UP instance to match the current traffic regime, achieving up to 23% energy savings during underutilization and 8.3% over full execution compared to static provisioning. The second scaling strategy follows the idea of the previous contribution, which implements latency-aware horizontal scaling by placing multiple CU-UP instances across the infrastructure and jointly minimizing energy consumption and midhaul latency, delivering up to 16.6% energy savings during underutilized regimes, 7.7% during non-peak periods, and 4.7% over the full execution. These two strategies are fundamentally complementary. Vertical scaling maximizes the energy efficiency but offers no latency improvement, while horizontal scaling reduces the end-to-end latency through geographic placement at the cost of lower energy savings.

However, the energy-aware orchestration of the CU-UP components addresses only part of the RAN energy picture. The DU and RU components represent the dominant share of the total RAN energy footprint, and managing their consumption at scale requires reliable power observability, which, in large-scale real-world deployments, makes external measurement hardware impractical. **RQ5** asked to what extent network KPIs could predict the power consumption of disaggregated RAN components, and how this relationship could support autonomous anomaly detection in online O-RAN deployments. **Chapter 7** addressed this by showing that the KPI telemetry already exposed by the O-RAN stack contains enough information to predict physical power consumption without any additional instrumentation. A hybrid XGBoost and CatBoost models trained on KPI measurements from a real 5G O-RAN deployment predicted O-RU power consumption with a MAPE of 1.02%, and O-DU power consumption with a MAPE of 1.22%. Based on these predictor models, a residual-

based anomaly detection framework was implemented that compares predicted power against hardware-reported measurements, achieving a 100% detection rate for manually injected anomalies and identifying genuine anomalies in the raw measurements.

Putting everything together, these five contributions provide a concrete and experimentally validated path toward 6G: heterogeneous RAT integration advances UC and IMC, the latency-aware CU-UP scaling targets IMC and HRLLC, ML-driven xApp control contributes to IAC, and energy-aware orchestration combined with KPI-driven power observability follow the Sustainability design direction.

8.2 Perspectives for Future Work

The findings of this thesis open several research questions for future research. Some of these directions represent the natural extensions of the individual contributions, while others reflect broader open challenges that emerged from the collective findings of the thesis.

8.2.1 Towards Full Multi-RAT Orchestration with Non-Terrestrial Networks

The Non-3GPP WiFi DU integration as presented in Chapters 3 and 4 demonstrated the real-world feasibility of the RAN-level heterogeneous in a two-RAT 5G system. The natural next step extension is to incorporate additional non-3GPP access technologies, such as NTN satellite links, into the same O-RAN control framework as described in Chapter 2. This is particularly crucial to the UC service category of 6G which explicitly defines the seamless integration of terrestrial and satellite into a unified 6G network. The already developed traffic steering xApp can be extended to manage routing decisions across these heterogeneous tiers, exploiting their complementary characteristics such as latency, throughput, reliability, and coverage profiles in order to dynamically select the most appropriate access path per user and per service class.

8.2.2 Hybrid and Cross-Layer RAN Scaling Strategies

Chapter 6 investigated the vertical and horizontal CU-UP scaling strategies independently. The experimental results demonstrated that these two scaling techniques have complementary

trade-offs in energy efficiency and latency compliance. The natural direction for future work is to design a hybrid scaling mechanism that dynamically selects or combines vertical and horizontal strategies to exploit the advantages of each.

Furthermore, this thesis focused only at the CU-UP component of the 5G RAN. Future work should also broaden the scope of the scaling framework to exploit the fully disaggregated RAN stack, including the DU component. A holistic cross-layer scaling framework that can simultaneously manage CU-UP and DU resource allocation can enable a system-wide optimization of throughput, latency, and energy consumption, providing stronger architectural foundations for the fully autonomous and self-managing RAN as 6G envisions.

8.2.3 Energy Management Across All O-RAN Components

Chapter 7 demonstrated that KPI-driven power prediction produces accurate and reliable results for the DU and RU components. The next immediate extension is to design a power prediction and anomaly detection mechanism for the CU component, enhancing the power consumption model developed in Chapter 6 by incorporating additional relevant KPIs that capture CU/CU-UP power behavior beyond incoming traffic load. This will enable a unified AI-based energy observability framework across all the disaggregated RAN components, enhancing the AIC service category of 6G. Additionally, deploying these KPI-driven power models as xApps at the Near-RT RIC would close the loop between energy prediction and resource control since predicted power deviations can autonomously trigger scaling or reconfiguration actions without relying on any physical measurement infrastructure, representing a concrete step toward a fully software-defined energy management system for 6G.

8.2.4 Foundation Models and Generative AI for RAN Management

SlAIces [8] demonstrated that LLMs can be effectively integrated into the configuration layer of the 5G experimentation phase, lowering the entry barrier for new researchers and accelerating deployment for experienced researchers. As already discussed in the Future Work of Chapter 6, a FM-driven xApp framework could unify the traffic forecasting and steering, scaling control, and power prediction capabilities developed across Chapters 4 - 7 into a single, continuously learning network intelligence layer. This unified FM, with access to the KPIs of all the O-RAN components (RU, DU, and CU/CU-UP), will be to able to identify simulatenously cross-layer dependencies and consequently perform optimization within the

RAN that the traditional controllers fundamentally are not so efficient. Furthermore, integrating such a framework with intent-based networking interfaces would enable operators to express high-level objectives in natural language. For e.g. an 6G administrator could simply state *"minimize energy during night hours while guaranteeing 10 ms latency for industrial users"* and the FM would translate it directly into xApp control parameters, without any manual configuration. Both directions align with the IAC service category of 6G and represent long-term but practically grounded research objectives.

8.2.5 Towards a Joint RAN and AI Workload Scheduling for AI-RAN

EELAS [7] demonstrated that cloud-native ML inference workloads can be efficiently scheduled across multiple heterogeneous infrastructure tiers such as edge, fog, and cloud, while jointly minimizing energy consumption and satisfying the set latency constraints. EELAS creates a natural extension motivated by the emerging AI-RAN vision [106], in which the RAN infrastructure becomes an AI compute platform colocated with the disaggregated radio functions.

A 6G AI-RAN deployment will simultaneously host three different workload classes: typical RAN workloads such as CU-UP and DU processing tasks, RAN AI-related workloads such as the xApp ML models developed across Chapters 4–7, and external AI workloads sourced from third-party latency-sensitive applications such as smart city analytics and industrial IoT processing. These classes have different requirements; for example, the RAN functions and their associated AI models (xApps) have very tight latency constraints, so placing an xApp model far from the RAN component could introduce control-loop latency. An extended EELAS framework that is aware of this affinity could jointly place and schedule all three workload classes under their heterogeneous latency, energy, and priority constraints, providing the scheduling substrate that makes AI-RAN practically deployable.

8.3 Lessons Learned

8.3.1 Real-World Testbed Experimentation as a Foundation for Applicability

In this thesis, all the proposed mechanisms were designed and evaluated on 5G testbeds under real-time operating conditions. This choice was not a methodological preference but instead a fundamental decision to establish the applicability of the proposed mechanisms. Although simulation environments are valuable for prototyping or large-scale exploration, they abstract away hardware-specific behaviors and timing dependencies that dominate real-world environments.

The experimental setup varied across the aforementioned contributions. In Chapters 3, and 4, the WiFi RAT operated under real RF conditions in an RF-isolated environment, offering an over-the-air validation of the heterogeneous RAN integration and the traffic steering mechanism. In Chapters 4 - 6, the 5G DU radio layer was emulated via the RF simulator, while all the other components, namely the CU/CU-UP, DU, near-RT RIC, and xApps operated on real hardware. Chapter 7 operated under real RF conditions for both the O-RU and O-DU, enabling the correlation of power consumption measurements with the reported KPIs from physical hardware, something that simulation environments cannot replicate.

Across all the contributions, the real-time operation of the RAN stack required strict timing constraints. The traffic steering xApp had to make the steering decisions before the traffic volumes changed and before each of the multi-RAT DUs became saturated. The traffic forecasting module presented in Chapters 4 - 5 was based on LSTM, while the remaining xApps relied on lightweight algorithms with low computational complexity in order to preserve the real-time nature of the control loop. This consideration became especially crucial in Chapter 6, where adding a traffic forecasting model at the beginning of the control pipeline required shorter inference times. The lightweight ML model of LightGBM was adopted instead of LSTM to avoid violating the real-time constraints of the xApp. Following the same principles, the use of lightweight ML models in Chapter 7 was a design choice, ensuring that the power prediction framework could operate in real-time and integrated in xApps to expose its calibrated metrics.

8.3.2 O-RAN as a Foundation for Self-Autonomous Networks

A key observation of this thesis is that O-RAN is not only a framework for deploying disaggregated RAN components, but also a necessary condition to enable autonomous network management. Before O-RAN, the control logic of the RAN was embedded in proprietary vendor systems, which made it very difficult to introduce external intelligence without strong cooperation from the vendor. O-RAN addresses this limitation by separating the intelligence layer from the RAN infrastructure. This intelligence is implemented via xApps and rApps running on the Near-RT and Non-RT RICs, and it interacts with the RAN through open, standardized interfaces.

This separation has an important consequence. A network that exposes its internal state through the O1, E2, and A1 interfaces can be monitored, analyzed and controlled by external AI-based applications, without requiring changes to the underlying RAN. The work presented in this thesis is based on this capability. Traffic steering, CU-UP scaling and power monitoring were implemented as independent xApps running on the Near-RT RIC. These xApps demonstrate that the O-RAN architecture can support multiple AI-driven control applications simultaneously, even across different RAN deployments.

Overall, this thesis provides a first practical step towards the vision of autonomous network management in 6G, where the RAN can continuously adapts to changing conditions with minimal human intervention.

8.3.3 Native AI Integration as a Design Principle for 6G RAN Management

The most important design principle emerging from this thesis is integration of AI not as an add-on, but instead as a core architectural component of the RAN control plane. This aligns directly with the 6G vision defined by ITU-R IMT-2030 [19], which identifies AI and ML as native components of future network architectures. As mentioned in Section 8.1, this thesis demonstrates that incorporating AI into the RAN control loop is not only feasible in real deployments but necessary to achieve the dynamic and autonomous network management of the 6G networks.

The traditional statistical models assume a stationary system behavior requiring manual tuning per deployment environment. In the RAN context, where the radio conditions are non-

stationary and continuously evolving, statistical models are fundamentally limited. On the contrary, AI models are able to learn directly from data and adapt the complexity of the problem. This is highlighted in Chapter 7, where the relationship between network KPIs and the physical power consumption of RAN components is inherently non-linear and environment-dependent. Tree-based models such as XGBoost and CatBoost captured these complex correlations, enabling a software-defined power consumption observability layer that can be used by other xApps in real-time.

8.4 Outlook

The contributions of this thesis were developed at a moment when the disaggregation of the RAN and the openness of its interfaces were transitioning from research concepts to operational reality. Looking ahead, the trajectory is clear: the networks of the 6G era will not be reactively reconfigured instead they will be self-learning and autonomously managed. Multi-RAT orchestration will extend beyond WiFi and 5G to encompass non-terrestrial networks, while foundation models and generative AI architectures will replace task-specific models with unified control agents that reason across multiple objectives and adapt to unseen network conditions. The O-RAN architecture provides the structural foundation for this vision. The work presented in this thesis is a simple step in that direction.

References

[1] Dimitris Kefalas, Nikos Makris, Serge Fdida, and Thanasis Korakis. Proactive scaling and energy optimization for 5g o-ran: Experimental insights. *IEEE Open Journal of the Communications Society*, 7:10626–10644, 2026.

[2] Dimitris Kefalas, Nikos Makris, Thanasis Korakis, and Serge Fdida. Design, Integration and Slicing of non-3GPP DUs in 5G Networks: a Hybrid 5G-DU Approach. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, Vancouver, Canada, May 2024. IEEE.

[3] Dimitris Kefalas, Nikos Makris, Thanasis Korakis, and Serge Fdida. Traffic steering for o-ran multi-rat 5g networks using ml-based demand forecasting. In *2025 IEEE Conference on Network Function Virtualization and Software-Defined Networking (NFV-SDN)*, pages 1–6, 2025.

[4] Dimitris Kefalas, Nikos Makris, Serge Fdida, and Thanasis Korakis. Enhancing 5g v-ran reliability via reactive cu-up scaling. In *2025 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, pages 292–297, 2025.

[5] Dimitris Kefalas, Nikos Makris, Serge Fdida, and Thanasis Korakis. On the feasibility of ran scaling for beyond 5g networks: Proactive cu-up scaling with ml demand forecasting. In *2025 IEEE 11th International Conference on Network Softwarization (NetSoft)*, pages 371–379, 2025.

[6] Dimitris Kefalas, Nikos Makris, Serge Fdida, and Thanasis Korakis. Prediction, anomaly detection and calibration of o-ran power consumption: An ai-driven approach using network kpis. In *2026 IEEE 12th International Conference on Network Softwarization (NetSoft)*, pages 497–502, 2026.

[7] Ilias Syrigos, Dimitris Kefalas, Nikos Makris, and Thanasis Korakis. Eelas: Energy efficient and latency aware scheduling of cloud-native ml workloads. In *2023 15th International Conference on COMmunication Systems NETworkS (COMSNETS)*, pages 819–824, 2023.

[8] Dimitris Kefalas, Sokratis Christakis, Serge Fdida, Nikos Makris, Ilias Syrigos, Virgilios Passas, and Thanasis Korakis. slaices: an llm chatbot for simplifying experiments with the slices-ri. In *2024 IFIP Networking Conference (IFIP Networking)*, pages 660–665, 2024.

[9] Caterina Leonelli, Dimitris Kefalas, Serge Fdida, Paolo Bellavista, and Thanasis Korakis. Dynamic resource allocation and energy optimization in 5g o-ran: Real-world insights and testbed evaluations. In *2025 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1019–1024, 2025.

[10] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. Empowering the 6G cellular architecture with open RAN. *IEEE Journal on Selected Areas in Communications*, 42(2):245–265, 2024.

[11] Haijian Sun, Yan Liu, Ahmed Al-Tahmeesschi, Avishek Nag, Milad Soleimanpour, Berk Canberk, Huseyin Arslan, and Hamed Ahmadi. Advancing 6G: Survey for explainable AI on communications and network slicing. *IEEE Open Journal of the Communications Society*, 6:1372–1412, 2025.

[12] NG-RAN; Architecture Description. Technical Report TS 38.401, 3rd Generation Partnership Project (3GPP), June 2018.

[13] Cheng-Xiang Wang, Xiaohu You, Xiqi Gao, Xiuming Zhu, Zi-sheng Li, Chuan Zhang, Haiming Wang, Yongming Huang, Yunfei Chen, Harald Haas, John S. Thompson, Erik G. Larsson, Marco Di Renzo, Wen Tong, Peiying Zhu, Xuemin Shen, H. Vincent Poor, and Lajos Hanzo. On the road to 6G: Visions, requirements, key technologies, and testbeds. *IEEE Communications Surveys & Tutorials*, 25(2):905–974, 2023.

[14] Abdikarim Mohamed Ibrahim, Rosdiadee Nordin, Angela Amphawan, and Muhammed Basheer Jasser. URLLC for 6G enabled industry 5.0: A taxonomy of architectures, cross layer techniques, and time critical applications. *arXiv preprint arXiv:2510.08080*, 2025.

[15] Afsoon Alidadi Shamsabadi, Animesh Yadav, Yasser Gadallah, and Halim Yanikomeroglu. Exploring the 6G potentials: Immersive, hyperreliable, and low-latency communication. *IEEE Vehicular Technology Magazine*, 20(1):74–82, 2025.

[16] Xuanyu Liang, Qiao Wang, Ahmed Al-Tahmeesschi, Swarna B. Chetty, David Grace, and Hamed Ahmadi. Energy consumption of machine learning enhanced open ran: A comprehensive review. *IEEE Access*, 12:81889–81910, 2024.

[17] Joonseok Lee, Hyunwoo Ryu, Youngkwon Kim, Yosub Kim, and Jungwoong Yi. TREES: Toward a real energy efficient system for 6G communications. *IEEE Wireless Communications*, 32(1):28–34, 2025.

[18] Line M. P. Larsen, Henrik L. Christiansen, Sarah Ruepp, and Michael S. Berger. Toward greener 5g and beyond radio access networks—a survey. *IEEE Open Journal of the Communications Society*, 4:768–797, 2023.

[19] ITU-R. Framework and overall objectives of the future development of IMT for 2030 and beyond. Recommendation ITU-R M.2160-0, International Telecommunication Union, 2023.

[20] Chilakala Sudhamani, Mardeni Roslee, Jun Jiat Tiang, Shahrina Khatun, and Md Ershadul Haque Chowdhury. A survey on 5G coverage improvement techniques: Issues and future challenges. *Sensors*, 23(4):2356, 2023.

[21] Navid Nikaein, Mahesh K. Marina, Saravana Manber, Christian Bonnet, Florian Kaltenberger, Dominique Nussbaum, and Riadh Ghaddab. OpenAirInterface: A flexible platform for 5G research. *ACM SIGCOMM Computer Communication Review*, 44(5):33–38, 2014.

[22] Robert Schmidt, Mikel Irazabal, and Navid Nikaein. FlexRIC: an SDK for next-generation SD-RANs. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '21, page 411–425, New York, NY, USA, 2021. Association for Computing Machinery.

[23] Qualcomm Atheros and Linux Kernel Community. ath10k: Qualcomm Atheros QCA988x Linux wireless driver. `https://wireless.wiki.kernel.org/en/users/drivers/ath10k`, 2024. Accessed: 2025-05-04.

[24] Jouni Malinen. hostapd: IEEE 802.11 AP, IEEE 802.1x/wpa/wpa2/eap/radius authenticator. `https://w1.fi/hostapd/`, 2024. Accessed: 2025-05-04.

[25] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.

[26] Brendan Burns, Brian Grant, David Oppenheimer, Eric Brewer, and John Wilkes. Borg, omega, and Kubernetes. *ACM Queue*, 14(1):70–93, 2016.

[27] Serge Fdida, Nikos Makris, Thanasis Korakis, Raffaele Bruno, Andrea Passarella, Panayiotis Andreou, Bartosz Belter, Cédric Crettaz, Walid Dabbous, Yuri Demchenko, and Raymond Knopp. Slices, a scientific instrument for the networking community. *Computer Communications*, 193:189–203, 2022.

[28] Darijo Raca, Dylan Leahy, Cormac J. Sreenan, and Jason J. Quinlan. Beyond throughput, the next generation: a 5G dataset with channel and context metrics. New York, NY, USA, 2020. Association for Computing Machinery.

[29] Yi Niu, Yong Li, Depeng Jin, Li Su, and Athanasios V. Vasilakos. Research on penetration loss of D-band millimeter wave for typical materials. *Sensors*, 22(19), 2022.

[30] Mohsen Nader Tehrani, Murat Uysal, and Halim Yanikomeroglu. The future is unlicensed: Coexistence in the unlicensed spectrum for 5G. *IEEE Communications Magazine*, 56(6):30–37, 2018.

[31] Dejun Deng, Jens Haas, and Athanasios V. Vasilakos. A survey on heterogeneous networks: Architecture, challenges and future directions. *IEEE Access*, 10:1–30, 2022.

[32] 3GPP. 3GPP TR 23.501 v15.3.0 (2018-09), 3rd Generation Partnership Project; Technical Specification; 5G; System Architecture for the 5G System , 2018.

[33] 3rd Generation Partnership Project (3GPP). 3GPP TS 23.501 V16.6.0 — System Architecture for the 5G System (5GS). 3GPP Technical Specification TS 23.501, ETSI / 3GPP, 2023.

[34] Katerina Pechlivanidou, Kostas Katsalis, Ioannis Igoumenos, Dimitrios Katsaros, Thanasis Korakis, and Leandros Tassiulas. Nitos testbed: A cloud based wireless experimentation facility. In *2014 26th International Teletraffic Congress (ITC)*, pages 1–6, 2014.

[35] A.K. Salkintzis. Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks. *IEEE Wireless Communications*, 11(3):50–61, 2004.

[36] Intel and Qualcomm. New WI proposal: LTE-WLAN Radio Level Integration and Interworking Enhancements, 2015.

[37] Alcatel-Lucent. Revised WID: LTE-WLAN Radio Level Integration and Interworking Enhancements, 2015.

[38] M. Luglio, M. Quadrini, C. Roseti, D. Verde, and F. Zampognaro. Performance evaluation of untrusted non-3GPP Access to a 5G Core Network via satellite. In *2022 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6, 2022.

[39] Mario Teixeira Lemes, Cristiano Bonato Both, Antonio Carlos De Oliveira Júnior, and Kleber Vieira Cardoso. A Tutorial on Trusted and Untrusted non-3GPP Accesses in 5G Systems - First Steps Towards a Unified Communications Infrastructure. *CoRR*, abs/2109.08976, 2021.

[40] Chengdong Yang, Tianyu Wang, Zhe Xiao, and Shaowei Wang. A Suggest-And-Improve RAN Slicing Strategy for Mobile Networks. In *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 748–752, 2022.

[41] Matteo Nerini and David Palma. 5g network slicing for wi-fi networks. In *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 633–637, 2021.

[42] Yiannis Yiakoumis, Kok-Kiong Yap, Sachin Katti, Guru Parulkar, and Nick McKeown. Slicing Home Networks. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Home Networks*, HomeNets '11, page 1–6, New York, NY, USA, 2011. Association for Computing Machinery.

[43] Matías Richart, Javier Baliosian, Joan Serrat, Juan-Luis Gorricho, and Ramón Agüero. Slicing With Guaranteed Quality of Service in WiFi Networks. *IEEE Transactions on Network and Service Management*, 17(3):1822–1837, 2020.

[44] Sven Zehl, Anatolij Zubow, and Adam Wolisz. Hotspot slicer: Slicing virtualized home Wi-Fi networks for air-time guarantee and traffic isolation. In *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–3, 2017.

[45] 3GPP. 3GPP TR 38.801 v14.0 (2017-03), 3rd Generation Partnership Project; Technical Report; Technical Specifications for Cloud Radio Access Network, 2017.

[46] Nikos Makris, Christos Zarafetas, Pavlos Basaras, Thanasis Korakis, Navid Nikaein, and Leandros Tassiulas. Cloud-Based Convergence of Heterogeneous RANs in 5G Disaggregated Architectures. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, 2018.

[47] R. De Vegt. 802.11ac usage models document. *IEEE 802.11-09/0161r2*, Mar 2009.

[48] 3rd Generation Partnership Project (3GPP). System architecture for the 5G System (5GS); Stage 2. Technical Report TS 23.501, 3GPP, 2020. Version 16.4.0.

[49] I. Chen, Shukri Abdallah, Feng Xie, and Venkatesh Ramaswamy. Using 5g to integrate heterogeneous networks with improved traffic engineering. In *MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM)*, pages 1192–1197, 2024.

[50] Alexander Rabitsch, Per Hurtig, Stefan Alfredsson, and Anna Brunstrom. Monty: A framework for latency-aware multi-flow atsss scheduling. In *2024 IEEE 49th Conference on Local Computer Networks (LCN)*, pages 1–9, 2024.

[51] Xueqing Zhou, Haiyuan Li, Anderson Bravalheri, Amin Emami, Reza Nejabati, Shuangyi Yan, and Dimitra Simeonidou. Drl-driven intelligent access traffic management for hybrid 5g-wifi multi-rat networks. pages 1–6, 2023.

[52] Fatemeh Kavehmadavani, Van-Dinh Nguyen, Thang X. Vu, and Symeon Chatzinotas. Intelligent traffic steering in beyond 5g open ran based on lstm traffic prediction. *IEEE Transactions on Wireless Communications*, 22(11):7727–7742, 2023.

[53] Rawlings Ntassah, Gian Michele Dell'Aera, and Fabrizio Granelli. xapp for traffic steering and load balancing in the O-RAN architecture. In *IEEE International Conference on Communications (ICC)*, pages 5259–5264, 2023.

[54] Andrea Lacava, Michele Polese, Rajarajan Sivaraj, Rahul Soundrarajan, Bhawani Shanker Bhati, Tarunjeet Singh, Tommaso Zugno, Francesca Cuomo, and Tommaso Melodia. Programmable and customized intelligence for traffic steering in 5G networks using open RAN architectures. *IEEE Transactions on Mobile Computing*, 23:2882–2897, 2023.

[55] Md Arafat Habib, Hao Zhou, Pedro Enrique Iturria-Rivera, Medhat Elsayed, Majid Bavand, Raimundas Gaigalas, Yigit Ozcan, and Melike Erol-Kantarci. Hierarchical reinforcement learning based traffic steering in multi-rat 5g deployments. In *ICC 2023 - IEEE International Conference on Communications*, pages 100–105, 2023.

[56] Ilias Chatzistefanidis, Nikos Makris, Virgilios Passas, and Thanasis Korakis. Ml-based traffic steering for heterogeneous ultra-dense beyond-5g networks. In *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2023.

[57] Lanlan Li and Tao Ye. Research on throughput prediction of 5g network based on lstm. *Intelligent and Converged Networks*, 3(2):217–227, 2022.

[58] Francisco Rau, Carlos Herranz, Iñaki Val, and Joaquin Perez. Lstm-based traffic prediction for 5g mobile network operator. In *2025 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–5, 2025.

[59] Xincai Fei, Fangming Liu, Hong Xu, and Hai Jin. Adaptive VNF Scaling and Flow Routing with Proactive Demand Prediction. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 486–494, 2018.

[60] Xiaoxi Zhang, Chuan Wu, Zongpeng Li, and Francis C.M. Lau. Proactive VNF provisioning with multi-timescale cloud resources: Fusing online learning and online optimization. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, 2017.

[61] Vasileios Zalokostas-Diplas, Nikos Makris, Virgilios Passas, and Thanasis Korakis. Experimental Evaluation of ML Models for Dynamic VNF Autoscaling. In *2022 IEEE Conference on Standards for Communications and Networking (CSCN)*, pages 157–162, 2022.

[62] Imad Alawe, Yassine Hadjadj-Aoul, Adlen Ksentinit, Philippe Bertin, Cesar Viho, and Davy Darche. An Efficient and Lightweight Load Forecasting for Proactive Scaling in 5G Mobile Networks. In *2018 IEEE Conference on Standards for Communications and Networking (CSCN)*, pages 1–6, 2018.

[63] Virgilios Passas, Nikos Makris, Yue Wang, Apostolos Apostolaras, Asterios Mpatziakas, Anastasios Drosou, Thanasis Korakis, and Dimitrios Tzovaras. Artificial intelligence for network function autoscaling in a cloud-native 5g network. *Computers and Electrical Engineering*, 103:108327, 2022.

[64] Sokratis Christakis, Nikos Makris, Thanasis Korakis, and Serge Fdida. On the Automated Scaling of User Plane Function for 5G: An Experimental Evaluation. In *2024 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, pages 979–984, 2024.

[65] Tulja Vamshi Kiran Buyakar, Anil Kumar Rangisetti, A Antony Franklin, and Bheemarjuna Reddy Tamma. Auto scaling of data plane VNFs in 5G networks. In *2017 13th International Conference on Network and Service Management (CNSM)*, pages 1–4, 2017.

[66] Davit Harutyunyan, Rasoul Behravesh, and Nina Slamnik-Kriještorac. Cost-efficient Placement and Scaling of 5G Core Network and MEC-enabled Application VNFs. In *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 241–249, 2021.

[67] Van-Giang Nguyen, Karl-Johan Grinnemo, Javid Taheri, Johan Forsman, Thang Le Duc, and Anna Brunstrom. On Auto-scaling and Load Balancing for User-plane Gateways in a Softwarized 5G Network. In *2021 17th International Conference on Network and Service Management (CNSM)*, pages 132–138, 2021.

[68] Abdirazak Ali Asir Rage, Ning Wang, and Rahim Tafazolli. Nfscaler: Ai-powered 5g-and-beyond network function scaler for qos assurance and energy efficiency. In *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*, pages 213–221, 2024.

[69] Akrit Mudvari, Nikos Makris, and Leandros Tassiulas. Exploring ML methods for Dynamic Scaling of beyond 5G Cloud-Native RANs. *ICC 2022 - IEEE International Conference on Communications*, pages 2284–2289, 2022.

[70] Khalid Ali and Manar Jammal. Ml-based dynamic scaling and traffic forecasting for 5g o-ran. In *2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 0444–0451, 2023.

[71] Sokratis Christakis, Theodoros Tsourdinis, Nikos Makris, Thanasis Korakis, and Serge Fdida. Evaluation of User Plane Function Implementations in Real-World 5G Networks. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, 2024.

[72] ETSI. 5G; Management and orchestration; 5G performance measurements (3GPP TS 28.552 version 16.9.0 Release 16). Technical Specification ETSI TS 128 552 V16.9.0, European Telecommunications Standards Institute, Sophia Antipolis, France, may 2021.

[73] Xingyi Cheng, Ruiqing Zhang, Jie Zhou, and Wei Xu. DeepTransport: Learning Spatial-Temporal Dependency for Traffic Condition Forecasting. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

[74] Hyeonjun Na, Yongjoo Shin, Dongwon Lee, and Joohyun Lee. LSTM-based throughput prediction for LTE networks. *ICT Express*, 9(2):247–252, 2023.

[75] Liheng Zhang and Yao Dong. 5G Cloud RAN on Intel Architecture. In *Proceedings of the 7th International Conference on Computer Science and Application Engineering*, CSAE '23, New York, NY, USA, 2023. Association for Computing Machinery.

[76] Charlotte Freitag and Mike Berners-Lee and Kelly Widdicks and Bran Knowles and Gordon S. Blair and Adrian Friday. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns*, 2(9):100340, 2021.

[77] Akrit Mudvari, Nikos Makris, and Leandros Tassiulas. Ml-driven scaling of 5g cloud-native rans. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2021.

[78] Imran Ashraf, Federico Boccardi, and Lester T. W. Ho. Sleep mode techniques for small cell deployments. *IEEE Communications Magazine*, 49, 2011.

[79] Md. Mehedi Hasan, Md. Haq, Md. Maruf Hossain, Junaed Tiash, and Sungoh Kwon. An energy-efficient small-cell operation algorithm for ultra-dense cellular networks. *IEEE Access*, 11:125245–125258, November 2023.

[80] Farinaz Kooshki, Ana García Armada, Md. Munjure Mowla, Adam Flizikowski, and Slawomir Pietrzyk. Energy-efficient sleep mode schemes for cell-less ran in 5g and beyond 5g networks. *IEEE Access*, 11:1432–1444, 2023.

[81] J. Xavier Salvat Lozano, Jose A. Ayala-Romero, Andres Garcia-Saavedra, and Xavier Costa-Perez. Kairos: Energy-efficient radio unit control for o-ran via advanced sleep modes. In *IEEE INFOCOM 2025 - IEEE Conference on Computer Communications*, pages 1–10, 2025.

[82] Derrick Wing Kwan Ng, Ernest S. Lo, and Robert Schober. Energy-efficient resource allocation in multi-cell ofdma systems with limited backhaul capacity. *IEEE Transactions on Wireless Communications*, 11(10):3618–3631, 2012.

[83] Emil Björnson, Luca Sanguinetti, Jakob Hoydis, and Mérouane Debbah. Designing multi-user MIMO for energy efficiency: When is massive MIMO the answer? In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, pages 242–247, Istanbul, Turkey, April 2014.

[84] N. K. Shankaranarayanan, Zhuohuan Li, Ivan Seskar, Prasanthi Maddala, Sarat Puthenpura, Alexandru Stancu, and Anurag Agarwal. Poet: A platform for o-ran energy efficiency testing. In *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, pages 1–5, 2024.

[85] Gianmarco Baldini, Raffaele Bolla, Roberto Bruschi, Alessandro Carrega, Franco Davoli, Chiara Lombardo, and Ramin Rabbani. Toward sustainable o-ran deployment:

An in-depth analysis of power consumption. *IEEE Transactions on Green Communications and Networking*, 9(2):429–444, 2025.

[86] Venkateswarlu Gudepu, Rajashekhar Reddy Tella, Carlo Centofanti, José Santos, Andrea Marotta, and Koteswararao Kondepu. Demonstrating the energy consumption of radio access networks in container clouds. In *NOMS 2024-2024 IEEE Network Operations and Management Symposium*, pages 1–3, 2024.

[87] Ujjwal Pawar, Bheemarjuna Reddy Tamma, and Franklin A Antony. Traffic-aware compute resource tuning for energy efficient cloud rans. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 01–06, 2021.

[88] Hiba Hojeij, Sahar Hoteit, Véronique Vèque, and Alexis I. Aravanis. Energy-efficient placement and association in disaggregated o-ran. In *Proceedings of the IEEE Conference on Network Softwarization and Management (CNSM)*, Bologna, Italy, Oct 2025. hal-05329935.

[89] Venkateswarlu Gudepu, Bhargav Chirumamilla, Rajashekhar Reddy Tella, Abhishek Bhattacharyya, Shubh Agarwal, Lavanya Malakalapalli, Carlo Centofanti, José Santos, and Koteswararao Kondepu. Earnest: Experimental analysis of ran energy with open-source software tools. In *2024 16th International Conference on COMmunication Systems NETworkS (COMSNETS)*, pages 1148–1153, 2024.

[90] Benoit Petit. scaphandre, 2023.

[91] Odin Ugedal and Rakesh Kumar. Mitigating unnecessary throttling in linux cfs bandwidth control. In *2022 IEEE 34th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pages 336–345, 2022.

[92] 3GPP. Management and orchestration; energy efficiency of 5G. Technical Specification TS 28.310, V18.4.0, 3rd Generation Partnership Project (3GPP), December 2023. Release 18.

[93] M. Schmidt and B. Singh. The balanced facility location problem: Complexity and heuristics. *INFORMS Journal on Computing*, 0(0), 2025.

[94] 3GPP. TS 38.473: NG-RAN; F1 Application Protocol (F1AP). Technical report, 3rd Generation Partnership Project (3GPP), 2021. Release 15.

[95] 3rd Generation Partnership Project (3GPP). NG-RAN; E1 Application Protocol (E1AP), 2019.

[96] Ana Larraáaga, Sandra Lagén, Josep M. Fàbrega, José Manuel Rivas-Moscoso, Juan Pedro Fernández-Palacios, Ioannis Tomkos, and Raül Muñoz. Fronthaul/midhaul networks: Capacity and latency requirements imposed by 6g disaggregated rans. *Comm. Mag.*, 63(5):86–93, May 2025.

[97] João Paulo S. H. Lima, George N. Katsaros, Rahim Tafazolli, and Konstantinos Nikitopoulos. A measurement-based power consumption model for networks with intelligent control (ric). *IEEE Open Journal of the Communications Society*, 6:2017–2031, 2025.

[98] Rasoul Behravesh, Daniel Coroama, Mastaneh Nikroo, Mohammadamin Gholami, Naresh Agrawala, and Rupendra Nath Mitra. Ai-driven traffic steering xapp for open ran: A cloud-native quantum-safe approach. In *2025 IEEE 11th International Conference on Network Softwarization (NetSoft)*, pages 49–54, 2025.

[99] Theodoros Tsourdinis, Ilias Chatzistefanidis, Nikos Makris, and Thanasis Korakis. Ai-driven service-aware real-time slicing for beyond 5g networks. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, 2022.

[100] Batuhan Akdogan, Ibrahim Hokelek, and Hakan Ali Cirpan. A federated learning framework for inter-cell handover optimization in o-ran. In *2025 16th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 1–5, 2025.

[101] Chuanhao Sun, Ujjwal Pawar, Molham Khoja, Xenofon Foukas, Mahesh K. Marina, and Bozidar Radunovic. Spotlight: Accurate, explainable and efficient anomaly detection for open ran. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '24, page 923–937, New York, NY, USA, 2024. Association for Computing Machinery.

[102] Godlove Suila Kuaban, Tadeusz Czachórski, Tulin Atmaca, and Piotr Czekalski. Power models for o-ran o-du planning and optimisation. In *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, NV, United States, January 2026. IEEE. HAL Id: hal-05386978.

[103] Jose A. Ayala-Romero, Ihtisham Khalid, Andres Garcia-Saavedra, Xavier Costa-Perez, and George Iosifidis. Experimental evaluation of power consumption in virtualized base stations. In *ICC 2021 - IEEE International Conference on Communications*, pages 1–6, 2021.

[104] Muhammad Qamar Usman, Cormac J. Sreenan, Marcin Dryjanski, and Aisling O'Driscoll. Power modeling of the o-ran o-ru application of advanced sleep modes for enhanced energy efficiency. In *2025 IEEE 22nd Consumer Communications Networking Conference (CCNC)*, pages 1–8, 2025.

[105] Makroz Ilyasse, Ahmed Nafidi, Ramón Gutiérrez-Sánchez, and Salaheddine Aourik. A machine learning approach for energy consumption in 5g networks. *ITU Journal on Future and Evolving Technologies*, 5:447–457, 12 2024.

[106] Naveed Ali Khan and Stefan Schmid. Ai-ran in 6g networks: State-of-the-art and challenges. *IEEE Open Journal of the Communications Society*, 5:294–311, 2024.