

ADAPTIVE CONNECTIVITY
MANAGEMENT IN
HETEROGENEOUS MOBILE
NETWORKS

DISSERTATION

A thesis submitted to the
FACULTY FOR COMPUTER SCIENCE, ELECTRICAL
ENGINEERING, AND MATHEMATICS
of
PADERBORN UNIVERSITY
in partial fulfillment of the requirements
for the degree of *Doctor rerum naturalium (Dr. rer. nat.)*

by

MARVIN ILLIAN

Paderborn
July 2026

SUPERVISOR:

Prof. Dr. Lin Wang

REVIEWERS:

Prof. Dr. Lin Wang

Prof. Dr. Marco Platzner

Prof. Dr. Antônio Augusto de Aragão Rocha

ORAL EXAMINATION COMMITTEE:

Prof. Dr. Lin Wang

Prof. Dr. Marco Platzner

Prof. Dr. Antônio Augusto de Aragão Rocha

Prof. Dr. Stefan Sauer

Dr. Ramin Khalili

DATE OF SUBMISSION:

July 2026

ABSTRACT

Modern cellular deployments offer heterogeneous radio-spectrum resources, yet connectivity management still relies on static, experience-driven parameters, creating two related inefficiencies: signal-strength-driven handover and band-switching decisions in CONNECTED mode ignore User Equipment (UE) energy consumption, while manually tuned idle-mode reselection parameters fail to adapt to dynamic load, mobility, and local heterogeneity. This dissertation investigates adaptive connectivity management across both regimes, aiming to replace static parameter choices with data-driven mechanisms that remain compatible with existing cellular procedures.

For CONNECTED mode, the thesis designs and implements an energy-aware band switching framework that jointly accounts for data-rate constraints, handover costs, and UE energy consumption, calibrated against commercial 5G power measurements. Simulations in a realistic deployment show that the framework reduces total UE energy consumption by up to 25.5 % relative to signal-strength baselines, establishing power consumption as a first-class mobility objective.

For IDLE mode, the thesis proposes CellPilot, a standards-compatible Reinforcement Learning (RL) framework that adapts area-wide cell reselection parameters from compact observations, balancing throughput, load balancing, and per-UE efficiency. The policy generalizes across unseen loads, topologies, temporal dynamics, and UE mobility. The single-agent control approach is extended to a cooperative multi-agent architecture where each cell agent controls local parameters using neighborhood-aware observations and compact global aggregates, preserving fixed action and observation spaces. Using local decisions and compact shared information, the distributed formulation remains scalable, outperforms heuristic baselines, and matches centralized performance, with median gains of up to 144 % in total network throughput, 24 % in load balancing, and 202 % in local UE throughputs.

Overall, the thesis demonstrates that data-driven, adaptive, and standards-compatible connectivity management improves UE energy efficiency and network performance in heterogeneous mobile networks.

ZUSAMMENFASSUNG

Moderne Mobilfunknetze bieten heterogene Ressourcen im Funkspektrum. Ihr Verbindungsmanagement bleibt jedoch weitgehend statisch und erfahrungsbasiert konfiguriert. Dies führt zu zwei Ineffizienzen: Im CONNECTED-Modus ignorieren signalstärkenbasierte Handover- und Bandwechselentscheidungen den Energieverbrauch der Endgeräte; im IDLE-Modus reagieren manuell gesetzte Zellauswahlparameter nur begrenzt auf Last, Mobilität und lokale Netzunterschiede.

Im CONNECTED-Modus entwirft und implementiert die Dissertation ein energiebewusstes Bandwechsel-Framework, das Datenraten, Handover-Kosten und den Energieverbrauch der Endgeräte gemeinsam berücksichtigt und mit kommerziellen 5G-Leistungsmessungen kalibriert wird. Simulationen in einer realistischen Netzkonfiguration zeigen, dass das Framework den gesamten UE-Energieverbrauch um bis zu 25,5 % gegenüber signalstärkenbasierten Referenzwerten reduziert und den Energieverbrauch als primäres Mobilitätsziel etabliert.

Für den IDLE-Modus führt die Dissertation *CellPilot* ein, ein standardkompatibles Framework für bestärkendes Lernen, das bereichsweite Zellauswahlparameter aus kompakten Beobachtungen anpasst und dabei Durchsatz, Lastverteilung sowie pro-UE-Effizienz ausbalanciert. Die Strategie generalisiert über unbekannte Lasten, Topologien, zeitliche Dynamiken und UE-Mobilität hinweg. Der Einzel-Agenten-Ansatz wird zu einer kooperativen Multi-Agenten-Architektur erweitert, in der jeder Zell-Agent lokale Parameter auf Basis nachbarschaftsbewusster Beobachtungen und kompakter globaler Aggregatwerte steuert, wobei feste Aktions- und Beobachtungsräume erhalten bleiben. Durch lokale Entscheidungen und kompakte geteilte Informationen bleibt die verteilte Formulierung skalierbar, übertrifft heuristische Referenzwerte und erreicht das Leistungsniveau des zentralisierten Ansatzes, wobei Mediansteigerungen von bis zu 144 % im Gesamtnetzdurchsatz, 24 % in der Lastverteilung und 202 % im lokalen UE-Durchsatz erreicht werden.

Insgesamt zeigt die Dissertation, dass datengetriebenes, adaptives und standardkompatibles Verbindungsmanagement die Energieeffizienz der Endgeräte und die Netzleistung in heterogenen Mobilfunknetzen verbessert.

FULL LIST OF AUTHOR PUBLICATIONS

The following works form the publication basis of this dissertation:

- [1] Marvin Illian, Ramin Khalili, Antonio A. de A. Rocha, and Lin Wang. “Learning-Based Cell (Re)Selection for Cellular Networks.” **Manuscript in preparation.** 2026.¹
- [2] Marvin Illian, Ramin Khalili, Antonio A. de A. Rocha, and Lin Wang. “Cooperative Multi-Agent Reinforcement Learning for Idle-Mode Cell (Re)Selection.” In: *Proceedings of the 51st IEEE Conference on Local Computer Networks (LCN)*. IEEE LCN 2026. Coimbra, Portugal: IEEE, Oct. 2026, pp. 1–13. DOI: [10.1109/LCN67947.2026.11660770](https://doi.org/10.1109/LCN67947.2026.11660770).
- [3] Marvin Illian, Ramin Khalili, Antonio A. de A. Rocha, and Lin Wang. “Cells on Autopilot: Adaptive Cell (Re)Selection via Reinforcement Learning.” In: *2026 24th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. **Best Paper Award.** Columbus, OH, USA: IEEE, June 2026, pp. 1–8. DOI: [10.23919/WiOpt71098.2026.11568224](https://doi.org/10.23919/WiOpt71098.2026.11568224).
- [4] Marvin Illian, Björn Luchterhandt, and Lin Wang. “Band Switching for Mobile Energy Optimization in 5G Networks and Beyond.” In: *Proceedings of the 20th Workshop on Mobility in the Evolving Internet Architecture*. MobiArch ’25. Hong Kong, China: Association for Computing Machinery, 2025, pp. 44–50. ISBN: 979-8-4007-1974-5. DOI: [10.1145/3737897.3767294](https://doi.org/10.1145/3737897.3767294).

Beyond the publications forming the basis of this dissertation, the author has contributed to a broader body of peer-reviewed research. The following publications are listed separately to document this wider publication record, including work from earlier research stages and complementary research directions.

¹ Manuscript to be submitted to IEEE Transactions on Networking. This journal article follows the fast-track publication opportunity associated with the WiOpt 2026 Best Paper Award.

- [5] Marvin Illian, Christopher Zinda, and Darius Schlangenotto. "Data Store Architectures: Balancing Functionality and Performance." In: *IEEE International Conference on Industrial Technology, ICIT 2024, Bristol, UK, March 25-27, 2024*. IEEE, 2024, pp. 1–6. DOI: [10.1109/ICIT58233.2024.10540811](https://doi.org/10.1109/ICIT58233.2024.10540811).
- [6] Arne Neumann, Marvin Illian, Tobias Harges, Lukas Martenvormfelde, Lukasz Wisniewski, and Jürgen Jasperneite. "An Architecture Concept for Short- and Long-term Resource Planning in the Industry 4.0 Environment." In: *18th IEEE International Conference on Factory Communication Systems, WFCS 2022, Pavia, Italy, April 27-29, 2022*. IEEE, 2022, pp. 1–4. DOI: [10.1109/WFCS53837.2022.9779161](https://doi.org/10.1109/WFCS53837.2022.9779161).
- [7] Marvin Illian, Simon Althoff, and Holger Karl. "A Process to Develop Lean Big-Data Platform Architectures for Industrial Manufacturing Contexts." In: *25th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2020, Vienna, Austria, September 8-11, 2020*. IEEE, 2020, pp. 993–996. DOI: [10.1109/ETFA46521.2020.9212006](https://doi.org/10.1109/ETFA46521.2020.9212006).²
- [8] Arnab Ghosh Chowdhury, Marvin Illian, Lukasz Wisniewski, and Jürgen Jasperneite. "An Approach for Data Pipeline with Distributed Query Engine for Industrial Applications." In: *25th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2020, Vienna, Austria, September 8-11, 2020*. IEEE, 2020, pp. 1005–1008. DOI: [10.1109/ETFA46521.2020.9212050](https://doi.org/10.1109/ETFA46521.2020.9212050).
- [9] Hadi Razzaghi Kouchaksaraei, Ashwin Prasad Shivarpatna Venkatesh, Amey Churi, Marvin Illian, and Holger Karl. "Dynamic Provisioning of Network Services on Heterogeneous Resources." In: *2020 European Conference on Networks and Communications, EuCNC 2020, Dubrovnik, Croatia, June 15-18, 2020*. IEEE, 2020, pp. 209–213. DOI: [10.1109/EUCNC48522.2020.9200912](https://doi.org/10.1109/EUCNC48522.2020.9200912).
- [10] Jörn Koepe et al. "Poster: First Performance Insights on Our Novel OFDM-based Vehicular VLC Prototype." In: *2018 IEEE Vehicular Networking Conference, VNC 2018*,

² This publication is based on work that originally formed part of the author's master's thesis. It is included in the publication list for completeness only; the work is not incorporated into the chapters of this dissertation and does not form part of its doctoral contributions.

- Taipei, Taiwan, December 5-7, 2018. IEEE, 2018, pp. 1–2. DOI: [10.1109/VNC.2018.8628322](https://doi.org/10.1109/VNC.2018.8628322).
- [11] Sevil Dräxler, Manuel Peuster, Marvin Illian, and Holger Karl. “Generating Resource and Performance Models for Service Function Chains: The Video Streaming Case.” In: *4th IEEE Conference on Network Softwarization and Workshops, NetSoft 2018, Montreal, QC, Canada, June 25-29, 2018*. IEEE, 2018, pp. 318–322. DOI: [10.1109/NETSOFT.2018.8460029](https://doi.org/10.1109/NETSOFT.2018.8460029).³
- [12] Sevil Dräxler, Manuel Peuster, Marvin Illian, and Holger Karl. “Towards Predicting Resource Demands and Performance of Distributed Cloud Services.” In: *KuVS-Fachgespräch Fog Computing 2018*. Ed. by Boris Koldehofe, Andreas Reinhardt, and Stefan Schulte. Technical Report. Vienna, Austria and Darmstadt, Germany, 2018, pp. 9–12. URL: <https://kuvs.de/fileadmin/FG/KUVS/1777.pdf>.⁴

COPYRIGHT NOTICES

In reference to IEEE copyrighted material which is used with permission in this thesis, IEEE does not endorse any of Paderborn University’s products or services. Internal or personal use of this material is permitted. If interested in reprinting or republishing IEEE copyrighted material for advertising or promotional purposes, or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a license from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

- ³ This publication originated from work conducted as part of the author’s bachelor’s thesis and was subsequently extended, revised, and published. It is listed here as part of the author’s broader publication record.
- ⁴ This publication is based on work that originally formed part of the author’s bachelor’s thesis. It is included in the publication list for completeness only; the work is not incorporated into the chapters of this dissertation and does not form part of its doctoral contributions.

ACKNOWLEDGMENTS

A doctoral journey is often described as a marathon. Mine included a pandemic, a change of supervisor, several project-related detours, and a final stretch that sometimes felt more like a sprint. Fortunately, it also included people who made the route possible and far more rewarding.

My path into research began with Professor Holger Karl. His courses sparked my interest in scientific work, and he later supervised my bachelor's and master's theses as well as the beginning of my doctorate. His precision, rigor, and careful reasoning continue to shape how I approach research. After his move to Potsdam, my work became increasingly self-directed, teaching me independence and persistence.

I am especially grateful to my supervisor and reviewer, Professor Lin Wang, for taking over my supervision in early 2024. Lin gave me substantial scientific freedom while always being available with calm advice and practical support. His trust and constructive approach created an environment in which research was both productive and enjoyable, helping to turn several years of experience and ideas into the coherent work presented here.

I thank Professor Stefan Sauer and the Software Innovation Lab for providing a stable institutional environment and project funding that enabled me to continue my research. I am grateful to Stefan for agreeing to serve on my doctoral committee. I also thank Professor Marco Platzner for reviewing this dissertation and serving on my doctoral committee.

My sincere thanks go to Dr. Ramin Khalili for serving on my doctoral committee and to Professor Antônio “Guto” Rocha for serving as a reviewer and committee member. I am grateful to both for their expertise, ideas, constructive criticism, and productive collaboration during the final phase of this work.

Finally, my deepest gratitude goes to my wife, Melany. She has accompanied me through this entire journey and shown remarkable patience with the flexible academic meanings of “almost finished” and “just one more thing”. Her trust and support gave me stability and reminded me that there is a life beyond papers, experiments, and deadlines. I am deeply grateful for the life we have built together and look forward to the next chapter of our lives.

CONTENTS

1	Introduction	1
1.1	Motivation and Context	1
1.2	Connectivity Management Across the User Equipment Lifecycle	3
1.3	Limitations of Conventional Approaches	3
1.4	Research Objectives and Core Contributions	5
1.5	Thesis Organization	6
 I Connectivity Management in Connected Mode		
2	Energy-Aware Band Switching in Connected Mode	11
2.1	Introduction	11
2.2	Background and Motivation	12
2.3	Energy-Aware Band Switching	13
2.4	A Case Study	25
2.5	Discussion and Conclusion	26
 II Connectivity Management in Idle Mode		
3	Idle Mode (Re)Selection in Cellular Networks	31
3.1	Idle-mode Cell (Re)Selection Procedure	31
3.2	Conventional Heuristic Baseline	34
3.3	Challenges of Global Load Balancing in Distributed Idle-Mode (Re)selection	36
4	Common Experimental Framework	37
4.1	Control Problem and Environment Model	37
4.2	Algorithm Selection and Refinement	40
4.3	Optimizations	46
4.4	Generalization	50
5	Adaptive Cell (Re)Selection via RL	53
5.1	Model	53
5.2	Evaluation	56
5.2.1	Evaluation Setup	56
5.2.2	Overall Performance and Generalizability	57
5.3	Conclusion	65
6	Cooperative Multi-Agent RL for Idle-Mode Cell (Re)Selection	69
6.1	Model	71
6.2	Learning	74
6.3	Evaluation	77

6.3.1	Evaluation Setup	79
6.3.2	Robustness Under Deployment-Oriented Constraints	81
6.3.3	Generalization Across Topology, Load, and Seed Diversity	86
6.3.4	Design Choices and Architecture Sensitivity	91
6.3.5	Analysis of UE QoE Improvements	93
6.4	Conclusion	95
7	Conclusion	99
7.1	Connected-Mode Synthesis	99
7.2	Idle-Mode Synthesis	100
7.3	Concluding Remarks	102
	Bibliography	105

LIST OF FIGURES

Figure 1.1	Traffic range where narrow-band operation is more energy-efficient	4
Figure 2.1	5G NR frame, subframe, and slot timing	14
Figure 2.2	DRX cycle with and without scheduled downlink data	16
Figure 2.3	UE power consumption over traffic load	20
Figure 2.4	Measured and modeled UE current in 5G SA and 5G NSA	22
Figure 2.5	Simulation area and base-station topology	26
Figure 2.6	Per-UE energy under low traffic load	27
Figure 2.7	Per-UE energy under high traffic load	27
Figure 3.1	Idle-mode cell reselection procedure	32
Figure 3.2	Per-UE throughput under two reselection configurations	34
Figure 4.1	Baseline simulation area in Petrópolis	40
Figure 4.2	Alternative simulation area in Petrópolis	41
Figure 4.3	Large simulation area in Petrópolis	41
Figure 4.4	Reward-baseline estimator mechanism	43
Figure 4.5	Continuous action parameter mapping	47
Figure 4.6	Policy network architecture	49
Figure 5.1	Single-agent training convergence	59
Figure 5.2	Single-agent performance under UE scaling	60
Figure 5.3	Single-agent cross-area generalization	61
Figure 5.4	Single-agent scalability to a large area	62
Figure 5.5	Single-agent performance under temporal dynamics	63
Figure 5.6	Single-agent ablation without curriculum learning	64

Figure 5.7	Single-agent convergence with 500 training seeds 64
Figure 5.8	Single-agent performance with increased training diversity 65
Figure 5.9	Single-agent robustness to UE mobility 66
Figure 5.10	Single-agent mobility performance with increased training diversity 66
Figure 6.1	Effect of baseline anchoring on multi-agent training 78
Figure 6.2	Effect of global information on multi-agent performance 81
Figure 6.3	Effect of UE mobility on multi-agent performance 82
Figure 6.4	Training dynamics under UE mobility 83
Figure 6.5	Effect of reduced control frequency on multi-agent performance 84
Figure 6.6	Training dynamics under reduced control frequency 85
Figure 6.7	Neighborhood graphs for geographic generalization 87
Figure 6.8	Multi-agent generalization to an alternative area 88
Figure 6.9	Multi-agent generalization across UE load levels 89
Figure 6.10	Effect of increased training diversity on multi-agent performance 89
Figure 6.11	Multi-agent training dynamics with 500 seeds 90
Figure 6.12	Effect of policy network size on multi-agent performance 91
Figure 6.13	Effect of baseline estimator window size on multi-agent performance 93
Figure 6.14	Effect of radio-quality histograms on multi-agent performance 94
Figure 6.15	Global UE throughput distribution 95
Figure 6.16	Neighborhood-level UE throughput distributions 97

LIST OF TABLES

Table 2.1	Input parameters for the power model	17
Table 2.2	Power model symbols	19
Table 2.3	Regression coefficients	21
Table 2.4	Index sets and parameters	23
Table 2.5	Decision variables	24
Table 3.1	Conventional heuristic reference values	35
Table 3.2	Cell (re)selection parameters, ranges, and purposes	35
Table 4.1	SNR-to-SE mapping	39
Table 4.2	REINFORCE training loop symbols	45
Table 4.3	Action space parameter ranges	46
Table 5.1	Single-agent observation components	54
Table 6.1	Comparison of single- and multi-agent formulation	71
Table 6.2	Multi-agent observation components	73
Table 6.3	Median gains across hidden-layer widths	92

LISTINGS

Listing 4.1	Episodic REINFORCE training loop	44
-------------	----------------------------------	----

ACRONYMS

5G NR	5G New Radio
5G NSA	5G Non-Standalone
5G SA	5G Standalone
A ₂ C	Advantage Actor-Critic
BWP	Bandwidth Part
CA	Carrier Aggregation
CSI	Channel State Information
DRX	Discontinuous Reception
ECDF	Empirical Cumulative Distribution Function
eNB	Evolved Node B
GRU	Gated Recurrent Unit
HOO	HandOver Parameter Optimization
KPI	Key Performance Indicator
LoS	Line of Sight
LSTM	Long Short-Term Memory
MILP	Mixed-Integer Linear Program
MIQCP	Mixed-Integer Quadratically Constrained Programming
ML	Machine Learning
MLP	Multi-Layer Perceptron
mmWave	Millimeter Wave
OSM	OpenStreetMap
PDCCH	Physical Downlink Control Channel
PDSCH	Physical Downlink Shared Channel
POMDP	Partially Observable Markov Decision Process
PRI	Parameter-Refresh Interval
PU	Power Unit
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
QoE	Quality of Experience

QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RF	Radio Frequency
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RRC	Radio Resource Control
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
SCS	Subcarrier Spacing
SDT	State Dwell Time
SE	Spectral Efficiency
SIB	System Information Block
SNR	Signal-to-Noise Ratio
SON	Self-Organizing Network
TDD	Time Division Duplex
UDR	Update-to-Dwell Ratio
UE	User Equipment

INTRODUCTION

1.1 MOTIVATION AND CONTEXT

The rapid deployment of 5G networks has substantially expanded the available radio spectrum, thereby sustaining the coexistence of low-band (sub-1 GHz), mid-band (1–6 GHz), and high-band (24 GHz and above) frequencies alongside existing 4G/LTE infrastructure [30]. Contemporary deployments have evolved beyond the limited frequency ranges of prior generations to support stable multi-band coexistence. Mobile devices typically access three to six bands, reaching up to eight in dense deployments, across 4G/LTE, 5G Non-Standalone (5G NSA), and 5G Standalone (5G SA). This multi-band access maintains spatial and temporal stability exceeding 95 % of operational time [29]. Operators leverage this consistent, heterogeneous radio environment to optimize capacity and coverage [29]. This spectrum heterogeneity enables adaptive connectivity management, allowing mobile devices to prioritize specific frequencies according to application demands, channel conditions, and network congestion. However, the availability of heterogeneous bands increases the complexity of mobility management, as operators must configure and coordinate network parameters across multiple Radio Access Technologies (RATs) to prevent service drops and optimize resource utilization across coverage areas.

Nevertheless, contemporary network operations realize only a limited portion of the potential offered by heterogeneous deployments. A fundamental challenge is the dual-sided nature of this limitation: from the operator perspective, static parameter configurations fail to exploit spectrum heterogeneity for resource efficiency, aggregate throughput, or fairness across users; from the device perspective, limited battery life constrains sustained performance and degrades overall Quality of Experience (QoE). Recent measurement studies indicate that the cellular radio system accounts for a substantial fraction of total smartphone power consumption, on the order of 30 % to 50 % in some measurements [40]. Operational parameters such as allocated bandwidth, transmission power, and traffic load significantly influence both energy demand and network utilization, yet exist-

ing mobility mechanisms rarely optimize for either energy efficiency or resource-aware performance. Consequently, achieving a balance between operator-level objectives—resource efficiency, total throughput, and fairness—and device-level constraints—per-UE throughput and power consumption—remains a critical challenge for operators in heterogeneous environments.

Connectivity management bridges spectrum potential and actual user experience, ensuring that theoretical capacity translates into tangible service quality. Network operators configure mobility parameters through static, experience-driven workflows, which are rarely adapted to dynamic network conditions. Mobility parameters control the critical aspects of network utilization and per-UE performance by governing how devices select and switch between cells. Optimizing coverage on one frequency may degrade throughput or increase energy consumption on another, due to significant variations in spectrum properties across bands. Manual parameter tuning and offline planning, characteristic of current operator practices, fail to keep pace with spatiotemporal traffic fluctuations, UE mobility patterns, or the rapid evolution of heterogeneous deployments. While the underlying network infrastructure remains constant, user experiences vary widely based on parameter configuration, leaving a substantial performance gap with static approaches.

Addressing this gap requires shifting from static, experience-based configuration to adaptive, data-driven management. Manual optimization is increasingly impractical due to the inherent complexity of heterogeneous networks, the energy constraints of modern UEs, and the dynamic nature of mobile traffic. This shift can be realized through two complementary paradigms: optimization-based approaches that solve precise problem instances using real-time measurements and empirically validated models, and learning-based approaches that infer optimal decisions from accumulated historical data. Together, these paradigms enable systems to respond to spatiotemporal patterns and adapt parameters in real time, facilitating autonomous connectivity management. The following sections examine how connectivity management mechanisms operate across the UE lifecycle, expose the limitations of conventional approaches, and outline the research objectives addressed in this thesis, thereby formalizing this motivation.

1.2 CONNECTIVITY MANAGEMENT ACROSS THE USER EQUIPMENT LIFECYCLE

Modern 5G architectures decompose the UE lifecycle into three distinct Radio Resource Control (RRC) states: RRC IDLE, RRC INACTIVE, and RRC CONNECTED. Connectivity management mechanisms governing mobile networks are tightly coupled to this lifecycle, operating across these states. Each state dictates a distinct mobility control paradigm that balances latency, power consumption, and signaling overhead. In RRC CONNECTED state, the UE maintains a dedicated radio context with the base station, enabling low-latency data transfer but resulting in the highest UE power consumption due to continuous resource reservation and network-controlled mobility through measurement reports and handover commands [13]. In RRC IDLE state, the UE performs autonomous cell selection and reselection based on broadcast thresholds, operating without an established RRC connection. Although this state maximizes power efficiency, it introduces higher wake-up latency and relies entirely on the UE's local measurements to select a camping cell [7]. As a 5G-specific innovation absent in LTE, RRC INACTIVE occupies an intermediate position between these states, retaining a lightweight context [28] in the anchor base station, enabling faster session resumption, UE-driven mobility, and extended periods of radio inactivity to conserve battery [15].

This thesis focuses on RRC CONNECTED and RRC IDLE modes, which represent the fundamental extremes of connectivity management from both user and network perspectives: RRC CONNECTED optimizes for user-perceived low latency and network-directed mobility control, whereas RRC IDLE prioritizes user energy efficiency and autonomous operation. RRC INACTIVE is excluded, as it acts merely as a transitional compromise between these states [7, 28].

1.3 LIMITATIONS OF CONVENTIONAL APPROACHES

Although both states implement distinct mobility procedures, both rely on parameter-configured control as the foundational mechanism for connectivity management. In CONNECTED mode, handover parameters (e.g., thresholds, offsets, hysteresis values, and events A3/A5) govern mobility by determining when and where the network triggers a band or cell switch [2]. In IDLE mode, reselection parameters (e.g., thresholds for higher

or lower priority frequencies, hysteresis values, and offset biases) determine how UEs camp on and switch between cells [7]. Despite real-time traffic dynamics, both mechanisms rely on parameter sets configured by operators, mostly based on experience, that operate independently of these conditions.

Connected-mode handover optimization faces a distinct limitation rooted in its optimization objective. Although 3GPP specifies a HandOver Parameter Optimization (HOO) Framework [5] to detect error scenarios such as too early or too late handovers, it does not prescribe specific algorithms for energy-aware optimization. Current mechanisms predominantly rely on Reference Signal Received Power (RSRP) measurements, prioritizing signal strength over actual UE energy consumption. As illustrated in Figure 1.1, this approach neglects the relationship between bandwidth allocation and power consumption: wider bandwidths consistently yield higher power consumption in low-to-medium traffic regimes, whereas high-power, narrow-bandwidth configurations can be more efficient within a certain traffic range. By optimizing for signal strength alone, current handover decisions neglect these energy implications, leaving substantial optimization potential in CONNECTED mode unexploited.

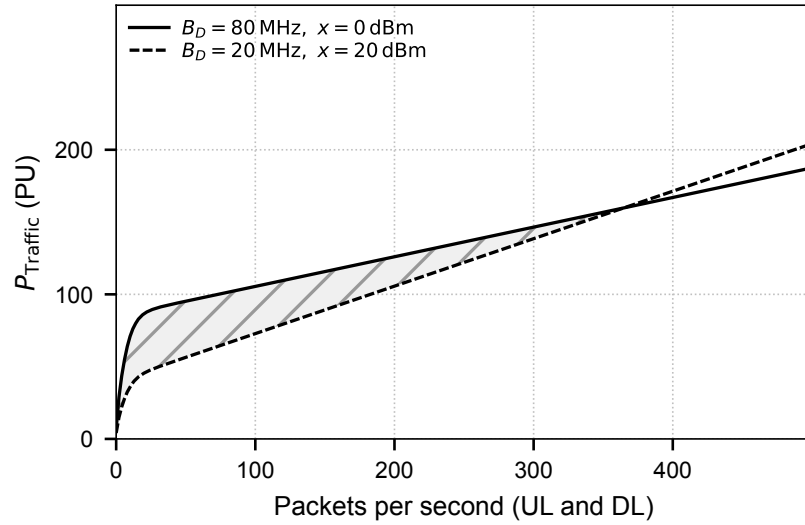


Figure 1.1: UE traffic power consumption P_{Traffic} for a wide, low-power configuration contrasted with a narrow, high-power configuration. A shaded interval denotes the specific traffic load where the 20 MHz BWP at 20 dBm achieves lower power draw than the 80 MHz BWP at 0 dBm, revealing that bandwidth scaling can supersede transmit power adjustments in governing UE energy consumption. Adapted from [37], © 2025 owner/author(s). Licensed to ACM.

Idle-mode cell (re)selection relies on static, operator-defined parameters that operate independently of real-time network conditions. These parameters are typically established through offline planning and experience-driven engineering, targeting aggregate Key Performance Indicators (KPIs) such as overall throughput or load-balancing behavior. Because these parameters are typically configured as a homogeneous set across a region within a distributed cellular network, modifying a single threshold can cascade through neighboring cells in unpredictable ways, indirectly altering load distribution and overall network utilization. This cascading effect, compounded by inherent spatiotemporal dynamics in traffic and user mobility, creates a high-dimensional, non-linear optimization space where the direct causal link between specific parameter adjustments and performance outcomes remains difficult to isolate. Consequently, operators typically deploy new parameter sets and monitor aggregate KPIs to assess impact, relying on heuristic, trial-and-error tuning that fails to scale with modern network complexity. Addressing this limitation requires shifting toward data-driven tools and automated optimization frameworks capable of navigating these coupled dynamics.

1.4 RESEARCH OBJECTIVES AND CORE CONTRIBUTIONS

This thesis addresses the limitations of static, experience-driven connectivity management by formulating and evaluating adaptive mechanisms that optimize UE energy efficiency in CONNECTED mode and enable autonomous, distributed control in IDLE mode. The primary research objective is to shift mobility management from static, operator-tuned configurations to data-driven, adaptive frameworks that respond to spatiotemporal network dynamics without requiring explicit mobility prediction or global state. To achieve this, we pursue three interconnected objectives:

1. designing and implementing an energy-aware band switching framework that models UE power consumption across bandwidth allocations and traffic loads, and formulates and solves runtime optimization as a Mixed-Integer Quadratically Constrained Programming (MIQCP) problem, aligning with the data-driven paradigm through empirical power measurements and real-time network state;

2. designing a single-agent Reinforcement Learning (RL)-based framework for adaptive idle-mode cell (re)selection that learns spatiotemporal patterns from limited aggregate observations; and
3. extending the formulation to a distributed, cell-specific multi-agent architecture that enables scalable, deployment-oriented control while preserving network-wide coordination.

1.5 THESIS ORGANIZATION

This thesis is structured into two parts, each addressing a distinct operational domain of connectivity management and advancing the transition from static configurations to adaptive, data-driven mobility frameworks through tailored optimization and learning approaches. Part I **CONNECTIVITY MANAGEMENT IN CONNECTED MODE** focuses on energy-aware band switching in active sessions. Chapter 2 begins by establishing the background and motivation for intelligent band switching, highlighting the suboptimal energy efficiency of **RSRP**-centric handovers. Subsequent sections develop a system-level **UE** power model calibrated against commercial 5G measurements, formulate energy-aware band switching as a **MIQCP** problem, and validate the framework through a case study over a realistic deployment, demonstrating energy optimization potential while preserving data rate demands. The final section discusses the transition from offline optimization to proactive, online intelligent band switching algorithms, outlining requirements for real-world deployment.

Part II **CONNECTIVITY MANAGEMENT IN IDLE MODE** addresses autonomous parameter adaptation in cellular reselection procedures. Chapter 3 reviews the 3GPP-specified idle-mode cell (re)selection process, exposing the limitations of static heuristic baselines and the coordination challenges inherent in distributed parameter configuration. Chapter 4 establishes a common experimental framework, defining the control problem, environment model, and **RL** optimizations, including history stacking, curriculum learning, and seed-specific baseline estimators. Chapter 5 presents a single-agent **RL** framework, CellPilot, evaluating convergence, **UE** scaling, cross-area generalization, and temporal dynamics. Chapter 6 extends the formulation to a cooperative multi-agent architecture, assigning one decision-making agent per cell and enabling distributed, cell-specific parameter control

through neighborhood-aware observations and shared global aggregates. The thesis concludes with a unified summary of findings and implications for self-optimizing Radio Access Networks (RANs).

Part I

CONNECTIVITY MANAGEMENT IN CONNECTED MODE

CONTENTS

2	Energy-Aware Band Switching in Connected Mode	11
2.1	Introduction	11
2.2	Background and Motivation	12
2.3	Energy-Aware Band Switching	13
2.4	A Case Study	25
2.5	Discussion and Conclusion	26

This part is based in part on material previously published as:
Marvin Illian, Björn Luchterhandt, and Lin Wang. “Band Switching for Mobile Energy Optimization in 5G Networks and Beyond.” In: *Proceedings of the 20th Workshop on Mobility in the Evolving Internet Architecture*. MobiArch ’25. Hong Kong, China: Association for Computing Machinery, 2025, pp. 44–50. ISBN: 979-8-4007-1974-5. DOI: [10.1145/3737897.3767294](https://doi.org/10.1145/3737897.3767294).

© 2025 Copyright held by the owner/author(s).

Publication rights licensed to ACM. The definitive Version of Record is available via DOI: [10.1145/3737897.3767294](https://doi.org/10.1145/3737897.3767294).

ENERGY-AWARE BAND SWITCHING IN CONNECTED MODE

This chapter focuses on energy-aware band switching during connected-mode operations. Unlike idle-mode selection, which governs initial camp-on behavior, connected-mode operations require continuous optimization of energy efficiency during active data transmission. We propose an energy-aware band switching mechanism that explicitly accounts for power consumption across varying bandwidths and traffic loads, and we evaluate its performance through a system-level power model and realistic network deployment.

2.1 INTRODUCTION

The rapid deployment of 5G networks has expanded the available radio spectrum significantly. German operators have announced 5G SA services [20] alongside their existing 4G/LTE and 5G NSA infrastructure. Unlike previous generations that operated over a limited range of frequencies, 5G supports a wide spectrum of bands, including sub-1 GHz, 1–6 GHz, and Millimeter Wave (mmWave). Future mobile networks will feature an even more diverse set of frequency bands [53]. This co-existence of heterogeneous bands enables mobile devices to select specific frequencies according to their respective use cases.

Despite this potential, current networks mostly select a default band for a UE upon entry, ignoring actual UE requirements and properties. Recent studies have explored intelligent band switching that considers UE requirements [29, 47]. For example, Hassan et al. [29] propose a hybrid architecture and algorithm that enables smart band decisions based on application QoE. Such designs could exploit heterogeneous bands to significantly boost mobile application performance. However, these existing works neglect UE power consumption, a major factor affecting user experience, in their band switching mechanisms. Recent measurement studies [26, 65] show that the smartphone cellular component dominates overall power consumption, highlighting the critical importance of cellular energy efficiency.

To address this limitation, we propose an energy-aware design for intelligent band switching in 5G networks and beyond. The rest of this chapter is structured as follows: Section 2.2 covers the background on 5G frequency bands, conventional band selection mechanisms, and the resulting energy efficiency challenges for UEs. In Section 2.3, we conduct comprehensive measurement studies on UE power consumption with respect to factors including data rate and allocated bandwidth. Based on these measurements, we build a power model for UEs. We then conduct a case study in Section 2.4 demonstrating how UE cellular energy efficiency could contribute to band switching decisions and yield associated gains. Section 2.5 discusses the remaining challenges in realizing a full-fledged energy-aware band switching design and provides a blueprint for future research directions.

2.2 BACKGROUND AND MOTIVATION

This section provides the background and motivates the adoption of smart band switching for optimized UE energy efficiency.

BAND SWITCHING IN 5G NETWORKS 5G networks utilize a wide range of frequency bands, including low-band (sub-1 GHz), mid-band (1–6 GHz), and high-band (24 GHz and above, also known as mmWave). Low-band frequencies offer wide coverage and good penetration but provide limited capacity. Mid-band frequencies provide a balanced trade-off between coverage and speed. High-band frequencies deliver ultra-high data rates but suffer from limited coverage and poor penetration, which requires dense infrastructure deployment. The complementary properties of these bands enable their integration to satisfy the diverse requirements of different environments. Current network deployments support an average of 4–8 available bands per region, a trend that remains temporally stable [29]. Furthermore, 5G coexists with 4G/LTE networks, offering a rich set of frequency bands to optimize mobile network performance. Mobile operators leverage this property by applying Carrier Aggregation (CA) to improve network bandwidth.

Despite the abundance of performance-varying bands, current mobile networks apply static band selection [29]. When a UE enters the network, it selects a default band based on idle-mode parameters configured by the base station, ignoring actual user requirements. Operating on this default band causes significant

time-varying network performance due to UE mobility. Band switching typically occurs during handover events, where the UE connection switches between cells that may operate on different frequency bands. However, existing band switching decisions are passive and sometimes function only as byproducts of other optimizations. Smart band switching, which continuously selects the frequency band that matches the best user experience, represents a large unexploited opportunity.

SUBOPTIMAL ENERGY EFFICIENCY. Alongside the suboptimal performance of passive band switching during handovers, we observe that the *energy efficiency* of the UE is often negatively impacted, directly affecting QoE. Current handover decisions predominantly rely on RSRP measurements, following the 3GPP specifications [14]. The primary optimization goal focuses on signal strength and transmission power requirements. These decisions typically neglect the power overhead associated with UE operation on different bandwidth allocations [55]. Cells may operate on different carrier frequencies and bandwidths, while the power consumption of a UE depends on the combination of its location and the allocated bandwidth.

Consider a UE running an application that requires a data rate of 100 Packets per Second (pps) across two neighboring cells. Cell A operates with a bandwidth of 80 MHz and requires 0 dBm transmission power for a reliable uplink. Cell B operates with 20 MHz bandwidth and requires 20 dBm transmission power. Both cells satisfy the data rate constraint. The legacy RSRP-based method selects Cell A to sustain the required 100 pps. Nevertheless, Cell B consumes less power due to energy savings achieved by operating on a narrower bandwidth. This claim is substantiated in Section 2.3 using power models based on 3GPP specifications and measurements conducted in a commercially deployed 5G network in Germany.

Existing works explore cellular energy consumption in handover signaling [62] and Discontinuous Reception (DRX) mechanisms [65]. Orthogonal to these prior works, we focus on the energy impact of band switching operations.

2.3 ENERGY-AWARE BAND SWITCHING

In this section, we build a model for UE power consumption and leverage it to optimize UE energy efficiency.

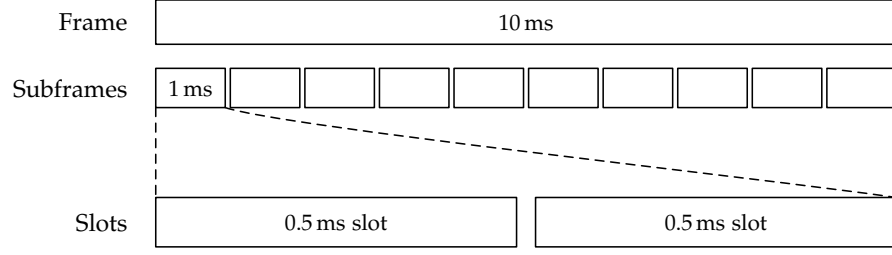


Figure 2.1: 5G NR timing hierarchy. A radio frame spans 10 ms and comprises ten 1 ms subframes [12]. For 30 kHz SCS, typically deployed in European and U.S. mid-band channels [25], each subframe contains two 0.5 ms slots.

POWER CONSUMPTION MODEL. In a joint initiative to reduce 5G New Radio (5G NR) UE power consumption, 3GPP and major hardware vendors published the “Study on UE Power Saving” in 2019 [9]. The resulting specification defines a power model that maps slot activities to slot-average power consumption for distinct UE power states, such as deep sleep, micro sleep, and PDSCH for downlink traffic. A slot serves as the fundamental 5G scheduling unit, where consecutive slots form a 1 ms subframe and a 10 ms radio frame [12]. The prevalent configuration in mid-band 5G deployments utilizes a slot duration of 0.5 ms [25]. European and U.S. 5G mid-band channels typically operate using 30 kHz Subcarrier Spacing (SCS) [25], the frequency separation between adjacent orthogonal subcarriers. Figure 2.1 illustrates the resulting timing hierarchy, showing how this configuration structures the radio frame, subframes, and slots.

The literature commonly assesses handovers via system-level simulations [50, 57, 58, 66]. These simulations lack millisecond-scale modeling and in-depth component representation. Consequently, a slot-level power model cannot easily integrate into system-level simulations. The developed power model must therefore predict average UE power consumption on a coarser time scale, motivating a system-level power model.

Based on the 3GPP study [9], we develop a theoretical system-level power model. Accordingly, we derive two analytical sub-models. The first quantifies the power overhead associated with each handover event, while the second models the continuous power consumption during CONNECTED data transmission. By averaging the power values of UE activities derived from system-level parameters, we predict the expected UE power consumption. Following the study [9], all power values are measured in

Power Units (PUs) relative to the UE's deep sleep power. This approach ensures a highly configurable and extensible system-level model. The handover duration is calculated by summing the estimated slot durations for synchronization and connection establishment yielding $T_{\text{ho}} = 117$ slots which corresponds to 58.5 ms at a 30 kHz SCS. Equation 2.1 defines the average UE handover power P_{ho} , which is parameterized by the UE transmission power x and the slot-average uplink power consumption P_{tx} [9].

$$\begin{aligned} P_{\text{ho}}(x \text{ [dBm]}) &= \frac{E_{\text{ho}}}{T_{\text{ho}}} = \frac{3.3 \cdot P_{\text{tx}}(x) + 6255.75}{117} \\ &\approx 0.064 \cdot 10^{\frac{x}{10}} + 60.46 \end{aligned} \quad (2.1)$$

The payload traffic power consumption P_{conn} depends on the UE transmit power, data rate, and the temporal distribution of packet arrivals. DRX is a 5G NR power-saving mechanism that limits downlink control-channel monitoring to periodic active windows separated by sleep intervals, with timing governed by the DRX cycle length, on-duration, and inactivity-timer [16]. The system-level model incorporates the parameters of power saving mechanisms like DRX and the Bandwidth Part (BWP) size. The subsequent discussion briefly elaborates on these mechanisms.

Figure 2.2 illustrates the DRX power-saving mechanism [17] across two cycles, demonstrating how traffic availability dictates UE activity states and power consumption. The Physical Downlink Control Channel (PDCCH) transmits scheduling assignments and control directives from the base station to the UE. Each cycle begins with a configurable on-duration during which the UE remains active to monitor the PDCCH for potential downlink traffic, resulting in elevated power draw. In the absence of scheduled downlink traffic, the on-duration expires, allowing the UE to enter a low-power sleep state for the remainder of the DRX period. The Physical Downlink Shared Channel (PDSCH) delivers scheduled user payload, system information, and broadcast messages from the base station to the UE. Conversely, when downlink traffic is detected and transmitted utilizing the PDSCH, an inactivity-timer is triggered, extending the active period to accommodate data transmission and keeping the UE awake beyond the initial on-duration. The figure highlights how DRX dynamically balances monitoring overhead with energy conservation based on actual traffic patterns.

The 5G NR bandwidth architecture follows a hierarchical structure, proceeding from the operating band to the cell channel

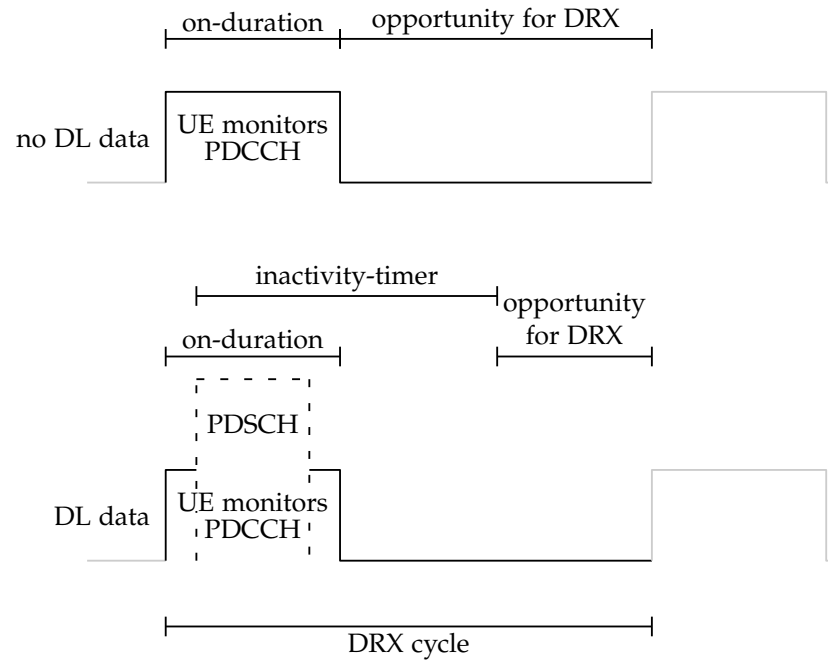


Figure 2.2: DRX cycle with and without scheduled downlink data. The upper panel, adapted from Figure 11-1 of [17], illustrates a DRX cycle without scheduled downlink data. The lower panel depicts the corresponding cycle under scheduled downlink traffic, where PDSCH reception activates the inactivity-timer and shortens the subsequent power saving opportunity for DRX.

Table 2.1: Input parameters for the power model

Parameter	Symbol	Unit
Mean downlink packet inter-arrival time	T_D	ms
Mean uplink packet inter-arrival time	T_U	ms
Average transmit power	x	dBm
DRX cycle length	T_C	ms
DRX inactivity-timer duration	T_I	ms
DRX on-duration	T_O	ms
Downlink BWP size	B_D	MHz

bandwidth, then to the UE-specific channel bandwidth, and finally to the configured BWP. Within this hierarchy, the BWP is formally defined as a subset of the UE-specific channel bandwidth. The UE power model captures bandwidth dependence via the downlink BWP size B_D . Band switching and Bandwidth Part switching are distinct mechanisms: the former changes the serving cell or carrier, while the latter modifies the active BWP within a single cell. In our system-level abstraction, however, we link them through B_D . We assume that the active downlink BWP spans the selected cell's channel, setting B_D equal to the cell bandwidth. Consequently, switching between narrow- and wide-band cells modifies the same bandwidth-dependent power term as standalone BWP switching, while additionally altering propagation conditions, cell load, and incurring handover energy overheads. This abstraction aligns with 3GPP UE power-saving studies that evaluate adaptation between full (100 MHz) and narrow (20 MHz) BWPs [9], as well as with European mid-band deployments in the 3.4 GHz to 3.8 GHz range, where Time Division Duplex (TDD) operation supports large contiguous channel bandwidths [21].

Table 2.1 specifies all input parameters required by the power model. We assume Poisson-distributed fixed-size packet arrivals, which aligns with the 3GPP "FTP 3" traffic model using 0.5 MB packets [3, 9]. The UE energy consumption over S slots is obtained by summing the slot-average power per activity plus the sleep transition power. The Physical Uplink Shared Channel (PUSCH) carries uplink user data, while the Physical Uplink Control Channel (PUCCH) enables the UE to report feedback, including acknowledgments, channel state information, and scheduling requests, back to the base station. We express the re-

sulting power model in Equation 2.2, where S_{PUSCH} denotes the number of uplink data-transmission slots and $S_{\text{PDCCH+PUCCH}}$ denotes the expected number of slots for uplink control signaling, including scheduling requests. S_{sleep} denotes the number of sleep slots. n_{sleep} denotes the average number of transitions into a sleep state within a single DRX cycle. S_{PDCCH} denotes the number of PDCCH-only slots. $S_{\text{PDCCH+PDSCH}}$ denotes the number of downlink slots with PDCCH and PDSCH [42].

We emphasize three specific inputs: the mean packet arrival times for downlink and uplink traffic T_D and T_U capturing traffic load, the UE's average transmit power x , and the downlink BWP size B_D . The DRX constants remain fixed. Equation 2.2 references slot counts S derived from the report [9] and power states P [9, 39], which are detailed in Table 2.2.

$$\begin{aligned}
 P_{\text{conn}}(T_D, T_U, T_C, T_O, T_I, x, B_D) \\
 = & \frac{1}{S} \left(P_{\text{deep}} \cdot S_{\text{sleep}} + P_{\text{deep}}^{\text{transition}} \cdot n_{\text{sleep}} \cdot \frac{T}{T_C} + \right. \\
 & S_{\text{BWP}}(B_D) (P_{r_{\text{CCH}}} (S_{\text{PDCCH}} + S_{\text{PDCCH+PUCCH}}) + \\
 & P_{\text{rx}} \cdot S_{\text{PDCCH+PDSCH}}) + \\
 & \left. P_{\text{short}}(x) \cdot S_{\text{PDCCH+PUCCH}} + P_{\text{tx}}(x) \cdot S_{\text{PUSCH}} \right) \quad (2.2)
 \end{aligned}$$

Applying the model to the example in Section 2.2, we predict power consumption across varying traffic loads. We vary the BWP B_D between 80 MHz and 20 MHz, alongside transmission power x values of 0, 20, and 23 dBm. Figure 2.3 visualizes the resulting tradeoffs. At 100 pps, the power consumption of the 80 MHz BWP with 0 dBm power exceeds that of the 20 MHz BWP with 23 dBm. This result demonstrates that cells requiring higher transmission power can still be more energy-efficient if they operate on a narrower bandwidth.

Summarizing Figure 2.3, we highlight four key findings. First, all curves exhibit an initial non-linear ramp-up followed by a nearly linear increase. Second, wider BWPs consistently yield higher power consumption than narrower ones, particularly in the low-to-medium traffic regime. Third, at a fixed BWP size, increasing the transmission power shifts the curves upward in parallel. Fourth, the curves intersect in the medium-to-high regime, indicating that high-power/narrow-bandwidth configurations can outperform low-power/wide-bandwidth configurations within a certain traffic range. These findings underscore the importance of incorporating data rate requirements into handover decisions for energy optimization. This approach enables

Table 2.2: Power model symbols

Symbol	Description	Value or definition	Reference
<i>Sleep</i>			
P_{deep}	Slot-average power assigned to deep-sleep slots.	$P_{\text{deep}} = 1 \text{ PU}$.	Table 18 in [9]
$P_{\text{deep}}^{\text{transition}}$	Additional overhead for entering and leaving deep sleep.	$P_{\text{deep}}^{\text{transition}} = 450$.	Table 19 in [9]
<i>Receive</i>			
P_{rCCH}	PDCCH monitoring power.	$P_{\text{rCCH}} = 100 \text{ PU}$ for PDCCH -only same-slot scheduling.	Table 18 in [9]
P_{rx}	Downlink receive power for slots with PDCCH and scheduled PDSCH reception.	$P_{\text{rx}} = 300 \text{ PU}$, corresponding to the PDCCH + PDSCH state.	Table 18 in [9]
$S_{\text{BWP}}(B_D)$	Scaling factor applied to downlink receive-power terms according to the active downlink BWP size B_D .	$S_{\text{BWP}}(B_D) = 0.4 + 0.6 \cdot \frac{B_D - 20}{80}$.	Table 21 in [9]
<i>Transmit</i>			
$P_{\text{short}}(x)$	Short PUCCH power.	$P_{\text{short}}(x) = 0.3 \cdot P_{\text{tx}}(x)$.	Table 21 in [9]
$P_{\text{tx}}(x)$	Slot-average uplink transmit power for PUSCH or long PUCCH at transmit power x in dBm.	$P_{\text{tx}}(x) \approx 250 + 2.267 \cdot (10^{x/10} - 1) \text{ PU}$.	Eq. 4 in [42]

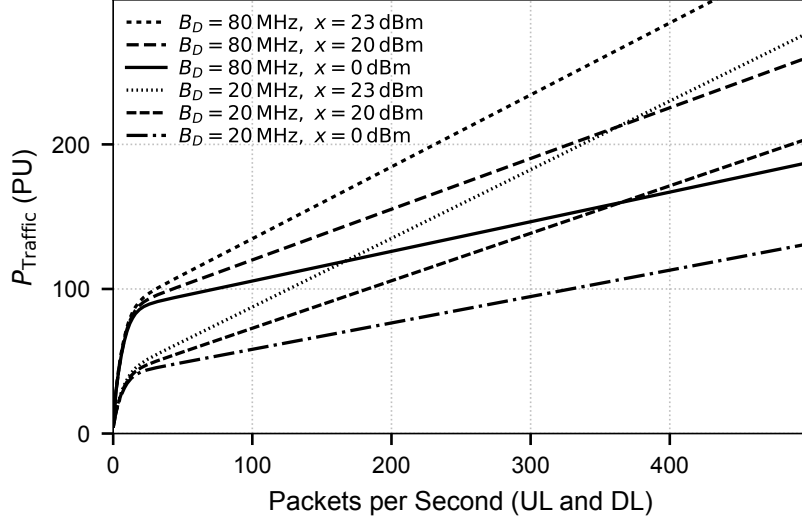


Figure 2.3: UE power consumption as a function of traffic load for various BWP and transmission-power configurations. Adapted from [37], © 2025 owner/author(s). Licensed to ACM.

fine-grained optimization based on the UE's actual load rather than relying on static signal strength criteria alone.

MODEL VALIDATION. To validate the model's accuracy, we compare it against measurements obtained in a commercial 5G deployment.¹ The real-world analytical model in Equation 2.2 requires calibration according to the specific UE and network configuration. To this end, we introduce per-summand coefficients obtained through measurement regressions shown in Table 2.3. Equation 2.3 presents a simplified closed-form expression to accelerate evaluation in large-scale simulations. Here, n_{sleep} denotes the average number of transitions into a sleep state per DRX cycle, and r_{sleep} represents the amount of sleep.

$$\begin{aligned}
 P_{\text{conn}}(T_D, T_U, T_C, T_O, T_I, x, B_D) \\
 \approx & \frac{1.473 \cdot 10^{\frac{x}{10}} - 0.375 \cdot B_D + 148.5}{T_U} \cdot c_1 \\
 & + \frac{0.75 \cdot B_D + 25}{T_D} \cdot c_2 + 0.75 B_D \cdot c_3 + 25 \cdot c_4 \quad (2.3) \\
 & - r_{\text{sleep}}(T_C, T_O, T_I, T_D, T_U) \cdot (0.75 \cdot B_D + 24) \cdot c_5 \\
 & + n_{\text{sleep}}(T_C, T_O, T_I, T_D, T_U) \cdot \frac{225}{T_C} \cdot c_6
 \end{aligned}$$

¹ The measurements were conducted using SIM cards provided free of charge by Deutsche Telekom AG. No proprietary information was shared, and the operator had no influence on the experimental setup or results.

Table 2.3: Regression coefficients

Coefficient	Value
c_1	0.229 916
c_2	0.182 944
c_3	2.801 657
c_4	1.037 651
c_5	1.786 270
c_6	43.200 216

We conducted two series of measurements to validate the model. The first series measured **5G SA** with a **BWP** of 90 MHz. The second series measured **5G NSA**, which aggregates a 20 MHz carrier and an LTE anchor, resulting in a total of 110 MHz. Each series was repeated five times and averaged to compensate for short-term fluctuations.

Figure 2.4 presents the measurements for both **5G SA** and **5G NSA**. The calibrated **5G SA** model aligns closely with the measurements, achieving a Root Mean Squared Error (**RMSE**) of 2.318 mA. We extend the model to **5G NSA** via an affine transformation of the **5G SA** measurements, optimizing a linear scaling factor and offset to minimize the error against the **5G NSA** measurements as shown in Equation 2.4.

$$\hat{y}_i^{\text{NSA}} = a \cdot y_i^{\text{SA}} + b \quad (2.4)$$

The optimal affine normalization yields a factor of $a = 1.16$ and an offset of $b = 31.27$ mA, resulting in an **RMSE** of 5.33 mA. Despite an aggregated bandwidth of 110 MHz, **5G NSA** operation typically only relies on the less energy-efficient LTE uplink (20 MHz) and employs **5G NR** solely for downlink transmissions (90 MHz). This configuration corresponds to a 22 % bandwidth increase. The factor a accounts for higher uplink transmission power compared to **5G SA**, as well as the traffic-dependent base-band processing load from parallel **RATs**. The offset b captures the traffic-independent power overhead from multiple Radio Frequency (**RF**) chains and control signaling, including Channel State Information (**CSI**) measurements and reporting. Applying the same affine normalization to the calibrated **5G SA** model yields a close fit to the **5G NSA** measurements with an **RMSE** of 6.04 mA.

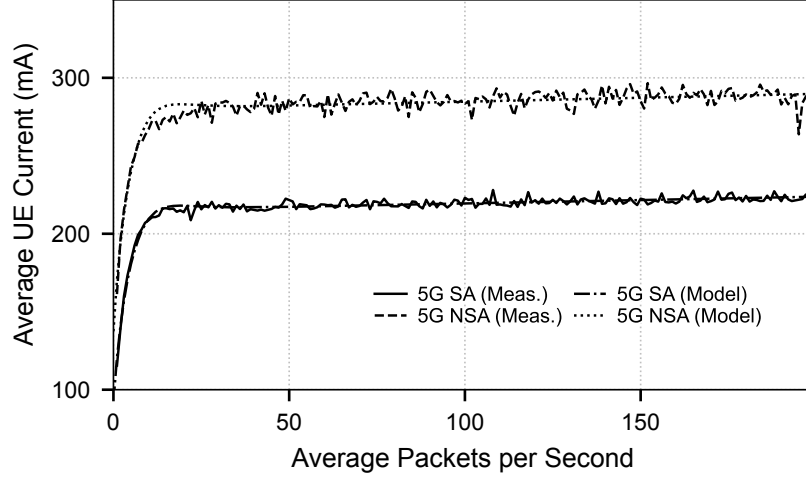


Figure 2.4: Measured and modeled average UE current consumption as a function of traffic load for 5G SA and 5G NSA. Both calibrated models closely follow the measurements, while 5G NSA consistently exhibits higher current as a result of the additional radio and signaling overhead inherent to non-standalone operation. Adapted from [37], © 2025 owner/author(s). Licensed to ACM.

Figure 2.4 also depicts the general effect of the DRX mechanism on UE power consumption. We map the PUs p to current in mA using the relation $I(p) = 1.55 \text{ mA} \cdot (p - 1) + 40 \text{ mA}$, derived from the measured deep sleep current and the UE's PDCCH power at a 90 MHz bandwidth. This effect is especially pronounced in low traffic regimes. After the initial ramp-up, both the measurements and the models demonstrate a linear power consumption trend. 5G SA consistently exhibits lower average currents than 5G NSA in the considered setup. The mean difference is 65 mA, corresponding to a mean 5G SA power reduction of 23 % compared to 5G NSA.

ENERGY OPTIMIZATION WITH BAND SWITCHING. The developed power models enable estimating UE energy consumption across various configurations. However, our goal is to facilitate runtime energy-aware band switching decisions. This requires selecting the most energy-efficient cell per UE and time slot while satisfying data rate requirements. These decisions are interdependent because cell load affects achievable data rates, which in turn alters energy consumption. To address this, we formulate a joint optimization problem that minimizes the total energy consumption across all UEs while balancing the distribution via variance minimization to ensure fairness among the UEs.

Table 2.4: Index sets and parameters

$\mathcal{U}$	Set of UEs, $ \mathcal{U} = U$
$\mathcal{C}$	Set of cells (may be in different bands), $ \mathcal{C} = C$
$\mathcal{T}$	Set of time slots, $ \mathcal{T} = T$
D_u	minimum data-rate demand of UE u [pps]
$\mathcal{N}$	Set of possible occupancies per cell, $\mathcal{N} = \{0, 1, \dots, U\}$
$E_{u,c,t}^{\text{conn}}$	Energy to stay connected on cell c during t [J]
$E_{u,c,t}^{\text{ho}}$	Handover energy if u switches into c at slot t [J]
$R_{u,c,t}^{(n)}$	Precomputed rate for UE u on cell c at slot t when exactly n UEs are attached [pps]
$E_{u,c,t}^{\text{conn},(n)}$	Precomputed connection energy when occupancy is n [J] on cell c at t

The resulting model captures both connection and handover energy costs subject to rate constraints and cell occupancy dynamics. The formulation constitutes a Mixed-Integer Linear Program (MILP) extended by a single convex-quadratic constraint (Equation 2.11), and is therefore a Mixed-Integer Quadratically Constrained Programming (MIQCP) problem.

To formulate the energy-aware handover and band switching problem, we utilize the index sets and parameters defined in Table 2.4. The model considers a finite time horizon of T discrete steps, a set of UEs $\mathcal{U}$, and a set of available cells $\mathcal{C}$ operating in possibly different bands. We introduce the binary decision variable $x_{u,c,t}$, which indicates whether UE u is connected to cell c at time t (see Table 2.5). An auxiliary binary variable $z_{u,c,t}$ denotes that UE u hands over to cell c at time t from a different cell, thereby incurring handover energy costs. Another auxiliary variable $y_{c,t,n}$ selects a concrete occupancy level $n \in \mathcal{N}$ per cell and slot, with $y_{c,t,n} = 1$ iff $\sum_{v \in \mathcal{U}} x_{v,c,t} = n$. Model constants, including rate tables $R_{u,c,t}^{(n)}$, connection energies $E_{u,c,t}^{\text{conn},(n)}$, and handover energies $E_{u,c,t}^{\text{ho}}$, are pre-computed via the power model described in Section 2.3.

The model enforces several constraints to ensure feasibility. Each UE is associated with exactly one cell per time step, expressed in Equation 2.5.

$$\sum_{c \in \mathcal{C}} x_{u,c,t} = 1 \quad \forall u \in \mathcal{U}, t \in \mathcal{T}. \quad (2.5)$$

Table 2.5: Decision variables

$x_{u,c,t} \in \{0,1\}$	1 if UE u is connected to cell c in slot t
$y_{c,t,n} \in \{0,1\}$	(Auxiliary occupancy) 1 if exactly n UEs are on cell c at slot t
$z_{u,c,t} \in \{0,1\}$	(Auxiliary handover) 1 if UE u hands over into c at slot t ($t \geq 2$)

The achieved data rate must meet the UE demand under the chosen occupancy, formulated in Equation 2.6.

$$\sum_{c \in \mathcal{C}} \sum_{n \in \mathcal{N}} R_{u,c,t}^{(n)} y_{c,t,n} x_{u,c,t} \geq D_u \quad \forall u \in \mathcal{U}, t \in \mathcal{T}. \quad (2.6)$$

Cell capacities are limited such that the total assigned bandwidth does not exceed the cell's available bandwidth, as formulated in Equation 2.7.

$$N_{c,t} = \sum_{u \in \mathcal{U}} x_{u,c,t} \quad \forall c \in \mathcal{C}, t \in \mathcal{T}. \quad (2.7)$$

Cell switching is captured via the logical difference between assignment variables across consecutive steps, as formalized in Equation 2.8.

$$\begin{aligned} z_{u,c,t} &\geq x_{u,c,t} - x_{u,c,t-1}, \\ z_{u,c,t} &\leq x_{u,c,t}, \\ z_{u,c,t} &\leq 1 - x_{u,c,t-1}, \end{aligned} \quad \forall u \in \mathcal{U}, c \in \mathcal{C}, t \in \mathcal{T} \setminus \{1\}. \quad (2.8)$$

The total energy per UE $u \in \mathcal{U}$, encompassing both traffic and handover components, is accounted for as E_u (Equation 2.9).

$$\begin{aligned} E_u &= \sum_{t=1}^T \sum_{c \in \mathcal{C}} \sum_{n \in \mathcal{N}} E_{u,c,t}^{\text{conn},(n)} y_{c,t,n} x_{u,c,t} \\ &\quad + \sum_{t=2}^T \sum_{c \in \mathcal{C}} E_{u,c,t}^{\text{ho}} z_{u,c,t} \quad \forall u \in \mathcal{U}. \end{aligned} \quad (2.9)$$

Our objective is to optimize the average UE energy consumption across all UEs, as formalized in Equation 2.10.

$$\mu = \frac{1}{U} \sum_{u \in \mathcal{U}} E_u. \quad (2.10)$$

To promote fairness, we include the standard deviation σ of per-UE energy in the objective, constrained by Equation 2.11.

$$\sigma^2 \geq \frac{1}{U} \sum_{u \in \mathcal{U}} (E_u - \mu)^2. \quad (2.11)$$

The overall objective minimizes $\mu + \sigma$ over the time horizon. The current formulation considers offline optimization to provide evidence of energy optimization potential. Real-world scenarios will require an online algorithm, as discussed in Section 2.5.

2.4 A CASE STUDY

To validate the proposed energy-aware band switching mechanism, we implement a case study using a Python simulation over a realistic 5G deployment. The assessment compares the optimization framework against a conventional RSRP-based baseline, measuring energy consumption and fairness under low and high traffic conditions.

We implement the optimization model introduced in Section 2.3 in a Python simulation. The simulation spans 10 steps, each representing 30 s of real time, for a total duration of 5 min. Consequently, handovers occur every 30 s. We simulate an area of 3.8 km² in Paderborn, Germany, using real base station positions to generate a realistic network topology. The topology consists of four cells, including two 5G SA and two 5G NSA base stations (see Figure 2.5). We assume the 5G NSA base stations operate using CA as measured in Section 2.3. All base stations are modeled to use a resource fair scheduler that assigns each connected UE an equal share of the total bandwidth. In the simulated area, we place 16 UEs with simulated mobility. One half of the UEs moves at a driving speed of 50 km/h, while the other half moves at a walking speed of 5 km/h. The movement trajectories of the UEs are random. We run the experiments with two load scenarios. In the *low load* scenario, each UE has a data rate demand of at least 10 Mbps. In the *high load* scenario, the data rate demand for each UE is increased to 30 Mbps.

As a baseline, we implement an RSRP-based mechanism that hands over UEs to the cell with the highest RSRP, ignoring power consumption. For comparison, we implement the optimization problem introduced in Section 2.3 in two variants: one minimizing μ only and one optimizing for $\mu + \sigma$. The resulting MIQCP is solved using the Gurobi Optimizer [27].

Across all 16 UEs, the energy-aware mechanism reduces the total energy budget compared to the baseline without compromising data rate demands. In the *low load* scenario, the energy budget is reduced by 25.5 % when minimizing μ and by 23.1 % when minimizing $\mu + \sigma$ combined (see Figure 2.6). Consequently, the standard deviation across UE energy consumption decreases

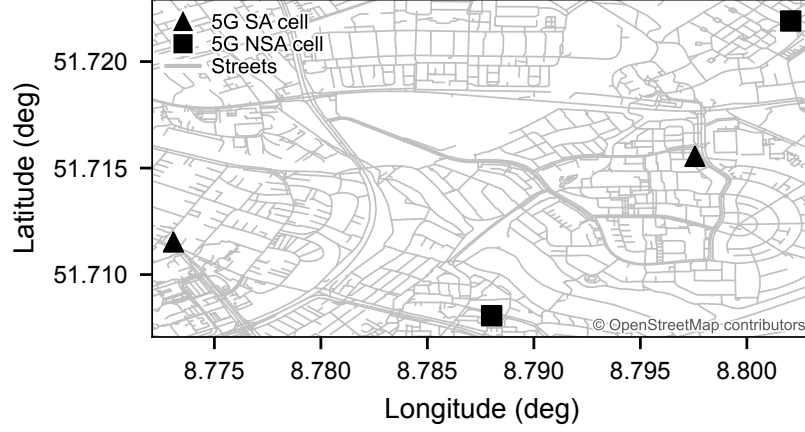


Figure 2.5: Map of the simulation area for the energy-aware band switching case study in Paderborn, Germany. It shows the OpenStreetMap [54] street layout and the real base-station locations used in the simulation, marking 5G SA cells and 5G NSA cells. Adapted from [37], © 2025 owner/author(s). Licensed to ACM.

by 32.2 % and 59.8 %, respectively. In the *high load* scenario, the energy budget is reduced by 8.9 % when minimizing μ and by 7.4 % when minimizing $\mu + \sigma$ (see Figure 2.7). The standard deviation across UE energy consumption increases by 4.9 % when minimizing μ and decreases by 20.3 % when minimizing $\mu + \sigma$. Note that the standard deviation is minimized over the entire simulation duration for all UEs. Therefore, it can occur that the standard deviation exceeds the baseline during individual steps even when $\mu + \sigma$ is minimized (e.g., time step 7 in Figure 2.6).

2.5 DISCUSSION AND CONCLUSION

This section synthesizes the technical contributions of this work and outlines the necessary extensions to transition from a validated proof of concept to a fully operational energy-aware band switching framework. We first address the architectural and algorithmic requirements for proactive switching and on-line decision-making under dynamic mobility constraints. The section concludes by summarizing the core findings and their implications for cellular network optimization.

PROACTIVE BAND SWITCHING. Our current design only makes band switching decisions at handover events, which aligns with the current 5G NR network architecture. Considering the abundance of frequency bands and their varying properties

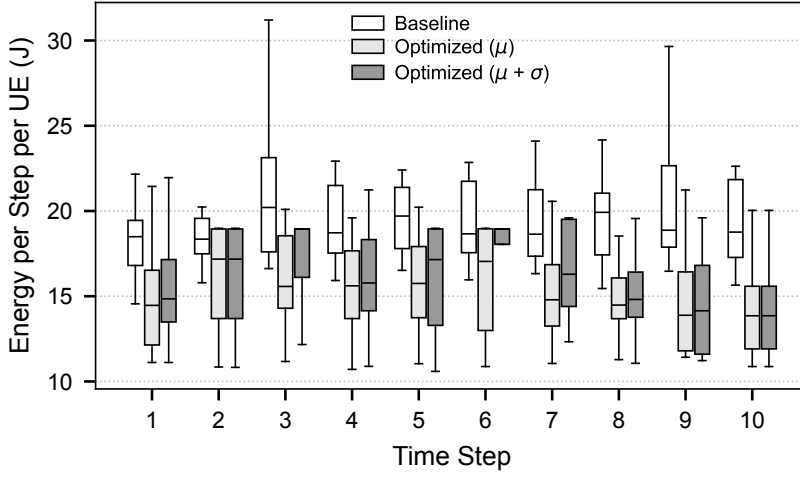


Figure 2.6: Distribution of per-UE energy consumption across simulation time steps in the low-load scenario with a 10 Mbps data rate demand per UE. Boxplots compare the RSRP-based baseline with the energy-aware optimization minimizing mean energy μ and the fairness-aware formulation minimizing $\mu + \sigma$. Adapted from [37], © 2025 owner/author(s). Licensed to ACM.

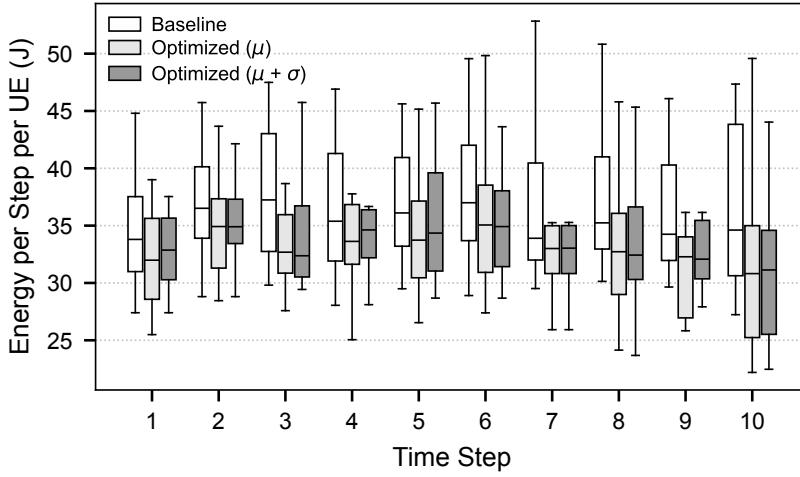


Figure 2.7: Distribution of per-UE energy consumption across simulation time steps in the high-load scenario with a 30 Mbps data rate demand per UE. Boxplots compare the RSRP-based baseline with the energy-aware optimization minimizing mean energy μ and the fairness-aware formulation minimizing $\mu + \sigma$. Adapted from [37], © 2025 owner/author(s). Licensed to ACM.

caused by constant UE mobility, it would be beneficial to perform *proactive* band switching whenever we can expect gains. However, this design requires the UE to continuously explore available bands, which incurs extra energy overheads. The exploration overhead must be accounted for in the power modeling and optimization. We consider this a promising direction for performance and energy efficiency optimization in increasingly heterogeneous mobile networks.

FROM OFFLINE OPTIMIZATION TO ONLINE LEARNING. Our current problem formulation is limited to the offline case, solving for a known sequence of future UE states. While our results demonstrate the potential for optimizing UE energy efficiency via band switching, real-world deployment is constrained by the unavailability of such future information at handover time. The optimization framework itself already constitutes a form of data-driven adaptation: it leverages a calibrated power model to produce energy-optimal band switching decisions, provided the future states are known. This creates a clear dependency: the optimization can operate online only if coupled with UE mobility models that predict future states from current measurements. We conjecture that Machine Learning (ML) methods can address this gap by learning implicit mobility patterns from historical data to inform the optimization’s future state inputs. RL aligns well with the requirement for continuous environment exploration and could learn switching policies directly from these learned mobility models. We identify the development of mobility-aware optimization frameworks as a key direction for bridging the gap between offline optimization and real-time decision-making.

CONCLUSION. We demonstrate that treating power consumption as a first-class metric fundamentally reshapes the band selection problem in cellular networks. We unify an analytically derived, empirically calibrated and validated UE power model with a problem formulation that considers UE power consumption during handovers and data transmission. Using simulations that reflect a real-world network topology, our method reduces overall UE energy consumption by up to 25.5 % relative to an RSRP-based baseline, while preserving fairness and meeting application data rate demands.

Part II

CONNECTIVITY MANAGEMENT IN IDLE MODE

CONTENTS

3	Idle Mode (Re)Selection in Cellular Networks	31
4	Common Experimental Framework	37
5	Adaptive Cell (Re)Selection via RL	53
6	Cooperative Multi-Agent RL for Idle-Mode Cell (Re)Selection	69

This part is based in part on material from the following publications and manuscript versions:

- Marvin Illian et al. “Cells on Autopilot: Adaptive Cell (Re)Selection via Reinforcement Learning.” In: *2026 24th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. 🏆 **Best Paper Award**. Columbus, OH, USA: IEEE, June 2026, pp. 1–8. DOI: [10.23919/WiOpt71098.2026.11568224](https://doi.org/10.23919/WiOpt71098.2026.11568224).

© IFIP, 2026. This material is reused under the author rights retained in the IFIP copyright agreement. The published version is available via IEEE Xplore: [10.23919/WiOpt71098.2026.11568224](https://doi.org/10.23919/WiOpt71098.2026.11568224).

- Marvin Illian et al. “Cooperative Multi-Agent Reinforcement Learning for Idle-Mode Cell (Re)Selection.” In: *Proceedings of the 51st IEEE Conference on Local Computer Networks (LCN)*. IEEE LCN 2026. Coimbra, Portugal: IEEE, Oct. 2026, pp. 1–13. DOI: [10.1109/LCN67947.2026.11660770](https://doi.org/10.1109/LCN67947.2026.11660770).

© 2026 IEEE. Reprinted with permission. The published version is available via IEEE Xplore: [10.1109/LCN67947.2026.11660770](https://doi.org/10.1109/LCN67947.2026.11660770).

- Marvin Illian et al. “Learning-Based Cell (Re)Selection for Cellular Networks.” 📄 **Manuscript in preparation**. 2026.²

² Manuscript to be submitted to IEEE Transactions on Networking. This journal article follows the fast-track publication opportunity associated with the WiOpt 2026 Best Paper Award.

IDLE MODE (RE)SELECTION IN CELLULAR NETWORKS

This chapter covers the 3GPP-specified idle-mode (re)selection process. The cell (re)selection process in mobile networks [7, 8, 43] is identical in 4G/LTE and 5G networks with only minor parameter differences.

3.1 IDLE-MODE CELL (RE)SELECTION PROCEDURE

Each cell in the network is powered by a base station operating on a specific carrier frequency and is identified by a unique cell ID [6]. Upon powering up and entering the network, the UE initiates the cell selection process, which is summarized in Figure 3.1 and covered in detail in the following. The selection criteria are determined by the RSRP measurements and the broadcast parameters provided by available cells. The initial selection establishes the camping point before the UE begins monitoring system information for subsequent idle-mode reselection.

SUITABILITY CONDITION VERIFICATION. Prior to considering a cell for (re)selection, the UE verifies that the signal satisfies a suitability condition to ensure adequate communication capability. This requirement is expressed as

$$S_{\text{rxlev},n}^{C_i} = \gamma_n^{C_i} - Q_{\text{rxlevmin}}^{C_i} > 0 \quad (3.1)$$

where $\gamma_n^{C_i}$ denotes the RSRP for UE n on cell C_i and $Q_{\text{rxlevmin}}^{C_i}$ specifies the minimum required RSRP for cell C_i as broadcast in the System Information Block (SIB). The resulting metric $S_{\text{rxlev},n}^{C_i}$ defines the cell (re)selection received level for UE n on cell C_i . Upon satisfying the suitability condition, the UE scans the available cells, selects the most appropriate candidate according to network-defined criteria, and camps on it. After camping, the UE continuously monitors the SIBs to remain informed about network parameter updates and suitable neighboring cells. Idle-mode (re)selection is subsequently governed by two complementary strategies: hierarchical cell (re)selection and equal priority cell (re)selection.

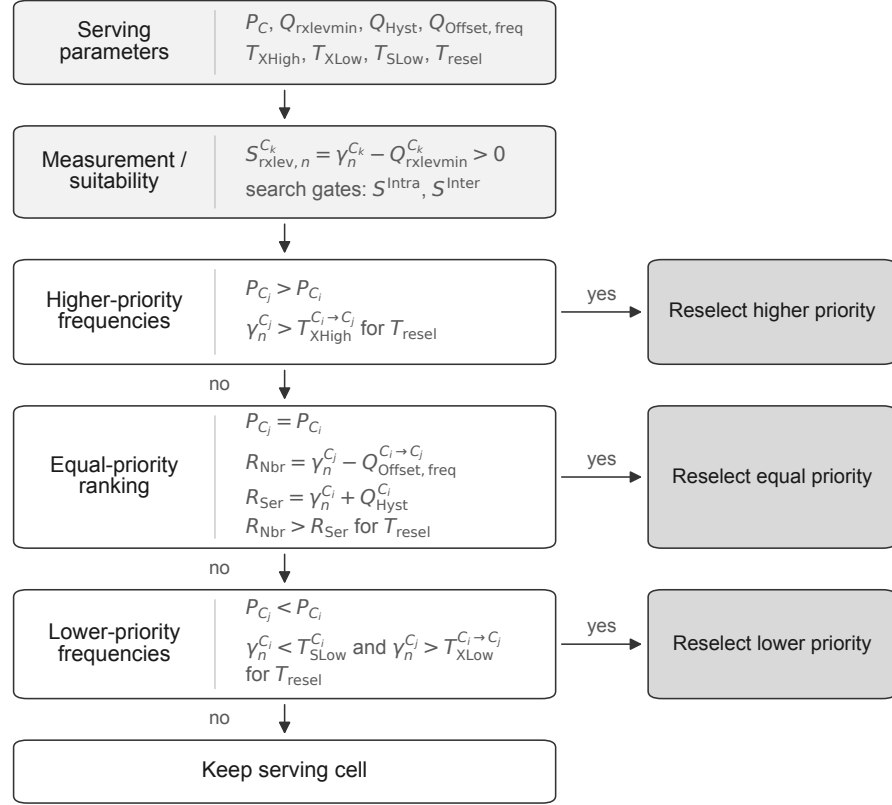


Figure 3.1: Idle-mode cell reselection procedure from serving-cell parameter evaluation to priority-specific reselection decisions. The procedure first verifies cell suitability and then evaluates higher-priority, equal-priority, and lower-priority candidates using their respective threshold, ranking, offset, hysteresis, and reselection-timer conditions.

HIERARCHICAL CELL (RE)SELECTION STRATEGY. In a hierarchical cell (re)selection strategy, the network assigns priorities P_{C_i} to cells to steer UEs toward optimal service and avoid highly congested cells, even when they exhibit higher signal strength. When the serving cell C_i holds a lower priority than a neighboring cell C_j , reselection criteria focus on the RSRP of C_j . The reselection process is initiated when the RSRP of the neighboring cell C_j exceeds the threshold $T_{XHigh}^{C_i \rightarrow C_j}$. This condition must be maintained continuously for the duration of the reselection timer T_{resel} :

$$\gamma_n^{C_j} > T_{XHigh}^{C_i \rightarrow C_j} \quad \text{for duration } T_{resel} \quad (3.2)$$

Enforcing this temporal requirement prevents frequent reselections caused by brief signal fluctuations. If multiple cells satisfy the criterion, the UE reselects the highest-ranked cell among the highest-priority frequencies. Reselection to lower-priority cells operates against the network's preferred hierarchy and involves multiple criteria. For inter-frequency reselection, the process begins when the serving cell's RSRP $S_{rxlev}^{C_i}$ falls below S^{Inter} , with an analogous intra-frequency threshold S^{Intra} applied. Lower-priority reselection is finally triggered if

$$\begin{aligned} \gamma_n^{C_i} &< T_{SLow}^{C_i} \quad \text{for duration } T_{resel} \quad \text{and} \\ \gamma_n^{C_j} &> T_{XLow}^{C_i \rightarrow C_j} \quad \text{for duration } T_{resel}. \end{aligned} \quad (3.3)$$

EQUAL PRIORITY CELL (RE)SELECTION STRATEGY. Equal priority cell (re)selection applies when the RSRP difference between cells falls below 3 dB. Under this condition, absolute signal levels become less critical, and the evaluation shifts to relative differences. This strategy directs users toward cells with stronger RSRP while enabling network operators to fine-tune traffic distribution through dynamic adjustment of the inter-frequency offset parameter (see Table 3.2). Increasing this offset discourages UEs from selecting specific neighboring cells, thereby balancing traffic according to prevailing congestion or Quality of Service (QoS) targets. The ranking mechanism evaluates cells based on RSRP and the offset parameter. The neighboring cell rank is defined as

$$R_{Nbr} = \gamma_n^{C_j} - Q_{Offset_{freq}}^{C_i \rightarrow C_j}, \quad (3.4)$$

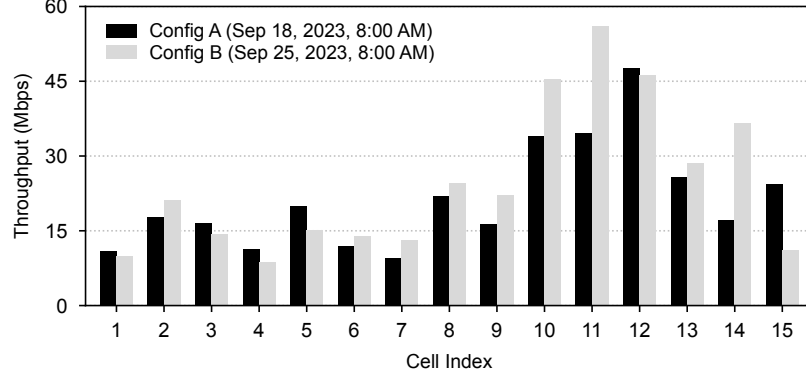


Figure 3.2: Per-UE downlink throughput across 15 cells in the Brazilian evaluation area under two cell reselection parameter configurations. The comparison demonstrates how operator-defined reselection configurations redistribute throughput among cells, motivating adaptive parameter optimization for idle-mode connectivity management. Adapted from [35], © 2026 IEEE.

while the serving cell rank incorporates a hysteresis value $Q_{\text{Hyst}}^{C_i}$ as

$$R_{\text{Ser}} = \gamma_n^{C_i} + Q_{\text{Hyst}}^{C_i}. \quad (3.5)$$

A timer T_{resel} again ensures robustness against brief RSRP fluctuations:

$$R_{\text{Nbr}} > R_{\text{Ser}} \quad \text{for duration} \quad T_{\text{resel}}. \quad (3.6)$$

3.2 CONVENTIONAL HEURISTIC BASELINE

The conventional heuristic baseline uses the network operator's parameter set shown in Table 3.1, derived directly from deployment data for Petrópolis, which also serves as the evaluation region in this study. Figure 3.2 compares two reselection configurations for this deployment and shows that parameter choices can substantially redistribute per-UE throughput across cells. We therefore adopt the area-specific Config-B as the heuristic baseline, ensuring that the RL approach is evaluated against a strong, tuned reference rather than a weak or generic configuration. Table 3.2 summarizes the complete set of cell (re)selection parameters, the ranges applied during simulation, and the specific purpose of each parameter within the reselection decision process.

Table 3.1: Conventional heuristic reference values

Symbol	Value
$Q_{\text{Hyst}}^{C_i}$	3 dB
$Q_{\text{Offset}_{\text{freq}}}^{C_i \rightarrow C_j}$	14 dB
$Q_{\text{rxlevmin}}^{C_i}$	−60 dBm
$T_{\text{XHigh}}^{C_i \rightarrow C_j}$	−56 dBm
$T_{\text{XLow}}^{C_i \rightarrow C_j}$	−58 dBm
$T_{\text{SLow}}^{C_i}$	−54 dBm

Table 3.2: Cell (re)selection parameters, ranges, and purposes

Symbol	Purpose
$T_{\text{XHigh}}^{C_i \rightarrow C_j}$	Threshold to reselect to higher priority cell
$T_{\text{XLow}}^{C_i \rightarrow C_j}$	Threshold to reselect to lower priority cell
$T_{\text{SLow}}^{C_i}$	Serving cell signal strength reselection threshold
$Q_{\text{Hyst}}^{C_i}$	Hysteresis to stabilize serving cell selection
$Q_{\text{Offset}_{\text{freq}}}^{C_i \rightarrow C_j}$	Offset biasing reselection between equal priority cells
S_{Intra}	Trigger threshold for intra-frequency measurements
S_{Inter}	Trigger threshold for inter-frequency measurements
$Q_{\text{rxlevmin}}^{C_i}$	Minimum required signal level for cell suitability
P_{C_i}	Cell priority for reselection decision-making

3.3 CHALLENGES OF GLOBAL LOAD BALANCING IN DISTRIBUTED IDLE-MODE (RE)SELECTION

Global load balancing presents a fundamental challenge in distributed idle-mode (re)selection because a UE evaluates neighboring cells using parameters broadcast by its current serving cell rather than by the target cell itself [7, Section 5.2.4.2]. Consequently, overloaded cells can effectively discourage UEs from camping on them, while underloaded cells lack the mechanism to directly attract UEs. The same target cell is therefore evaluated differently by UEs camped on different serving cells, meaning that a UE on one cell may reselect to a target due to favorable parameters while another UE on a different cell may not. The load on a target cell thus depends on the parameter configurations of multiple neighboring cells, creating a coordination problem across source cells. As a result, the final load on any single cell emerges from the combined parameter settings of several serving cells rather than a single local decision, rendering load balancing inherently a coordination problem across agents.

Local UE gain is relatively more directly attributable to local serving-cell actions because it depends on the radio conditions and reselection outcomes of the UEs currently camped on that cell. A beneficial parameter change can therefore more immediately improve the experience of those UEs without requiring complex coordination across multiple cells. Similarly, global throughput can benefit from numerous local improvements in signal conditions and cell selection, even without perfectly balancing the load across the entire network.

In contrast, load-balancing gain is a more indirect and spatially diffuse objective because parameter changes must propagate through serving-cell-specific UE evaluations before altering the aggregate occupancy of neighboring cells. This propagation also makes credit assignment harder in a multi-agent setting, since an individual agent observes its own UEs and serving-cell performance more directly than the eventual network-wide load consequences of its actions. This dynamic aligns with the neighborhood-only ablation (see Figure 6.2), where removing global observations and reward terms preserves strong throughput and UE gains but sharply reduces load-balancing gain. Thus, throughput and local UE gains can arise from local serving-cell improvements, whereas load balancing requires indirect, serving-cell-specific steering and cross-cell coordination.

COMMON EXPERIMENTAL FRAMEWORK

This chapter establishes the shared experimental framework and RL pipeline used throughout this part. Section 4.1 defines the control problem, environment dynamics, and network topology models; Sections 4.2 and 4.3 detail the policy-gradient algorithm and optimization techniques, while Section 4.4 discusses generalization strategies employed for parameter adaptation. The resulting framework provides a reproducible and robust foundation for the subsequent idle-mode experiments.

4.1 CONTROL PROBLEM AND ENVIRONMENT MODEL

This section defines the control problem and the environment model for idle-mode cell (re)selection. We formulate the control problem as a discrete-time decision process in which one or multiple agents periodically observe the available network state and update reselection parameters. We assume a discrete-time system where steps are decomposed into two distinct step types: Simulation steps update the environment; for example, UEs may move or switch their state (IDLE/CONNECTED). During agent steps, the agent retrieves the current observation, takes an action, and receives a reward. This decoupling allows the agent to operate at a different frequency than the underlying simulation. While every agent step necessarily triggers a corresponding simulation step, a simulation step does not always trigger an agent step. When the agent operates at a lower frequency, it receives observations only every n -th simulation step and cannot observe the intermediate states directly; *history stacking* can partially mitigate this loss of temporal information by retaining previous observations (see Section 4.3).

Because key dynamics remain hidden, including UE mobility, traffic states, per-UE signal conditions, and the actions of other agents, the control task is modeled as a Partially Observable Markov Decision Process (POMDP). To ensure reproducibility, we assign a fixed random seed to each episode, which governs the stochastic processes for UE modes and positions. The environment seed generates a unique scenario by combining cell topology and UE placement, while each UE receives a de-

rived seed for its stochastic processes controlling mobility and transitions between IDLE and CONNECTED modes throughout the episode. Episode lengths can be varied during training (see *curriculum learning* in Section 4.3).

UE MODELING AND MOBILITY. UEs are placed at realistic street positions. To ensure a high degree of realism, UEs are positioned according to real-world street layouts retrieved from OpenStreetMap (OSM). The simulator supports configurable UE mobility; when enabled, UEs move along streets at an average speed of 30 km/h, providing a realistic representation of urban movement patterns. While the absolute duration of a simulation step in the static case is arbitrary and defined solely by the relationship between mode transitions and update frequencies, this velocity provides a physical connection to the real world, anchoring our discrete time steps to a one-second interval. Even when disabled, the agent is trained on a variety of scenarios with unique UE positions, using a training set of 100 unique scenarios, each containing 500 UEs with randomly seeded traffic and positions. Validation is conducted against 100 unseen scenarios to assess the generalizability of the learned policy across different UE distributions and traffic patterns. The reselection timer T_{resel} , which defines the duration a candidate cell must sustain its signal strength before triggering a reselection, is fixed to one simulation step.

CHANNEL, PROPAGATION, AND RESOURCE ALLOCATION MODEL. We implement a custom Python-based simulator that realizes the cellular network model, providing an experimental foundation enabling learning-based parameter adaptation. The model is based on Line of Sight (LoS) geometry and a sectorized antenna pattern with a 120° sector width. Cell positions and azimuths are extracted from real-world operator deployment data, while the beamwidth is inferred from the spatial layout and visual cell coverage. We model the antenna gain as 10 dB within the main lobe [18] and 0 dB in all other directions, as a simplifying assumption. The path-loss model follows a free-space model at GHz frequencies. Thermal noise is modeled as $-174 \text{ dB mW Hz}^{-1}$ following the Johnson-Nyquist model [1]. In this noise-limited environment, link quality is expressed by the Signal-to-Noise Ratio (SNR), and Spectral Efficiency (SE) is computed via a nearest-neighbor lookup based on the SNR-to-SE mapping shown in Table 4.1 [10, Tab. 5.2.2.1-2], [46]. Resource al-

Table 4.1: SNR-to-SE mapping

Modulation	Code rate	SE	SNR (dB)
QPSK	78/1024	0.152	−9.478
QPSK	120/1024	0.234	−6.658
QPSK	193/1024	0.377	−4.098
QPSK	308/1024	0.602	−1.798
QPSK	449/1024	0.877	0.399
QPSK	602/1024	1.176	2.424
16QAM	378/1024	1.477	4.489
16QAM	490/1024	1.914	6.367
16QAM	616/1024	2.406	8.456
64QAM	466/1024	2.730	10.266
64QAM	567/1024	3.322	12.218
64QAM	666/1024	3.902	14.122
64QAM	772/1024	4.523	15.849
64QAM	873/1024	5.115	17.786
64QAM	948/1024	5.555	19.809

location is performed via a resource-fair scheduler that allocates equal bandwidth to all active UEs in each cell using an iterative water-filling algorithm.

GEOGRAPHICAL AREA AND CELL TOPOLOGY. Experiments are conducted over multiple distinct geographical areas characterized by unique topologies. Cell tower information, including locations, azimuth, frequency, bandwidth, and reselection parameters, is extracted from a real-world dataset provided by a major mobile operator in Brazil. Where reselection parameters were missing, we inferred representative values from public sources [24, 51, 60]. This approach of deriving default parameters from multiple authoritative sources, including vendor specifications and industry best practices, when complete operator configuration datasets are unavailable is validated by industry guidelines [52]. All evaluated areas are located in Petrópolis, Brazil, and include three distinct configurations:

- The **Baseline area** consists of two towers and 15 cells (see Figure 4.1), covering 2.36 km².

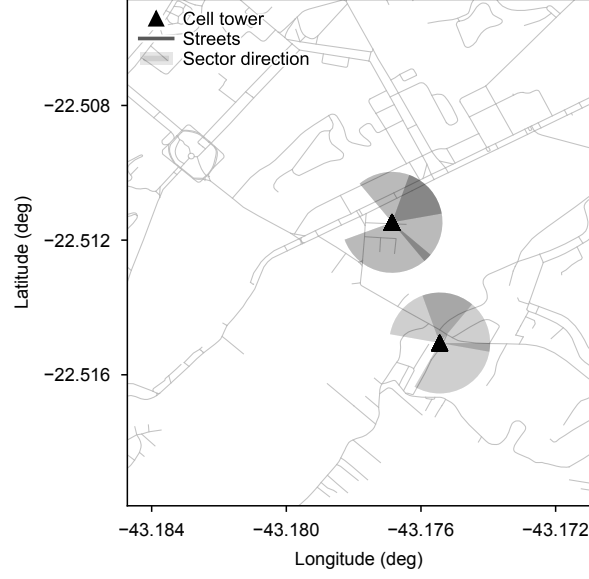


Figure 4.1: Baseline simulation area in Petrópolis, Brazil, for the idle-mode cell reselection experiments. The OpenStreetMap [54] street layout is combined with the actual operator cell-tower locations and sector orientations, covering two towers and 15 cells across 2.36 km^2 .

- The **Alternative area** features the same number of towers, cells, and area size (see Figure 4.2) as the baseline but is geographically distinct to assess policy transferability across different deployment configurations.
- The **Large area** covers a significantly larger area of 6.10 km^2 with six towers and 48 cells (see Figure 4.3), which we use to evaluate the scalability of the proposed learning formulation.

4.2 ALGORITHM SELECTION AND REFINEMENT

RL methods are typically categorized into value-based and policy-based paradigms. Value-based approaches estimate action values and select actions by comparing these estimates, which inherently requires a discrete action space and becomes computationally intractable as the action set grows [61]. As discussed in Chapter 3, in 3GPP-compliant cell (re)selection, the network does not perform explicit UE-to-cell assignments; instead, control is achieved by adjusting continuous reselection parameters (e.g., hysteresis, thresholds, offsets), necessitating a continuous action space. Policy-based methods parameterize

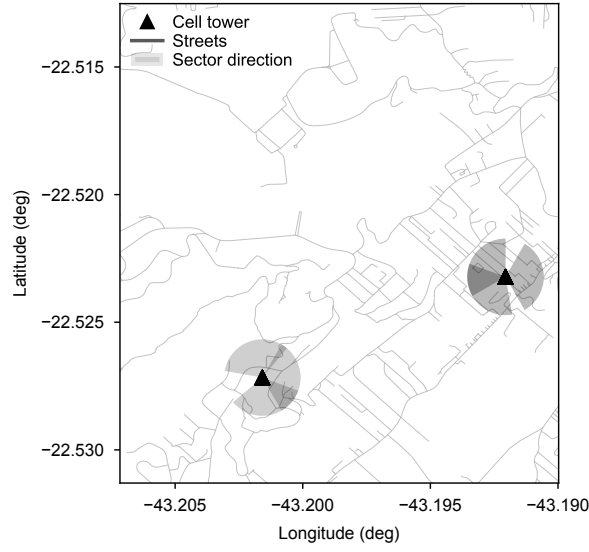


Figure 4.2: Alternative simulation area in Petrópolis, Brazil, used to evaluate transferability across deployment areas. The area maintains the same scale as the baseline scenario, covering two towers and 15 cells across 2.36 km^2 , but features geographically distinct tower placements and sector orientations. Map data from OpenStreetMap [54].

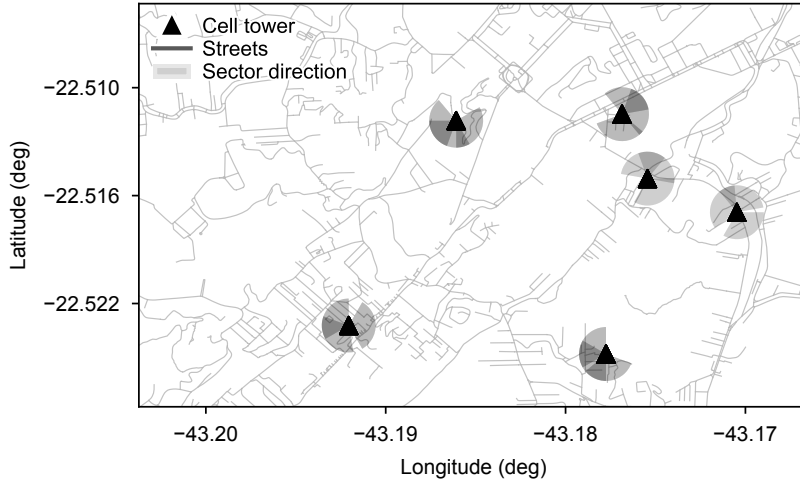


Figure 4.3: Large simulation area in Petrópolis, Brazil, used to evaluate the scalability of the learning formulation. The OpenStreetMap [54] street layout is combined with the actual operator cell-tower locations and sector orientations, encompassing six towers and 48 cells across 6.10 km^2 . Adapted from [33], © 2026 IFIP.

the action distribution directly, naturally accommodating continuous or high-dimensional action spaces without enumerating or approximating action values.

Actor-critic methods such as Advantage Actor-Critic (A2C) extend this class by combining a value function for credit assignment with a parameterized policy, thereby supporting continuous action spaces while retaining explicit policy parameterization [61]. We therefore experimented with the popular A2C method [48] but observed no convergence, likely due to problem-intrinsic dynamics. We hypothesize that varying seeds produce different total throughputs and therefore different temporal reward profiles creating non-stationary targets for the value function. As a result, the critic network struggles to converge while attempting to fit a mixture of unrelated target values. This issue is further discussed in [44], which highlights the inherent limitations of value-based approaches in scheduling environments. A similar underlying effect may also be present in our environment.

To resolve this, we adopt the vanilla REINFORCE policy-gradient algorithm [64] with seed-specific moving-average baseline references [45], which estimate the expected reward as a relative performance difference compared to a historical reference. Since REINFORCE computes gradients directly from sampled rewards without a shared critic, it avoids gradient interference. For each seed s , the baseline $B_s(t)$ is computed as a moving average over the last w episodes at step t . Because episodes sharing a seed maintain consistent topology and traffic, the baseline serves as an unbiased estimator of the expected return. We compute these baselines separately for each optimization goal, using the conventional heuristic performance as a reference if the estimator is still unpopulated. Figure 4.4 depicts the data flow among the baseline estimator, its keyed metric history, the conventional-heuristic reference, and the resulting baseline-normalized reward term.

REINFORCE ALGORITHM. The episodic REINFORCE training loop [61, Ch. 13.3] implements a policy-gradient approach tailored to continuous reselection parameters, as detailed in Listing 4.1. We parameterize the action distribution as a Gaussian $\mathcal{N}(\mu_t, \sigma_t^2)$ conditioned on the current observation s_t , with the output layer producing $\mu_t \in [0, 1]$ and $\sigma_t \in [0, 0.1]$ to bound the normalized action mean and exploration variance. Each sampled action a_t is clipped to the valid interval $[0, 1]$ to ensure feasible

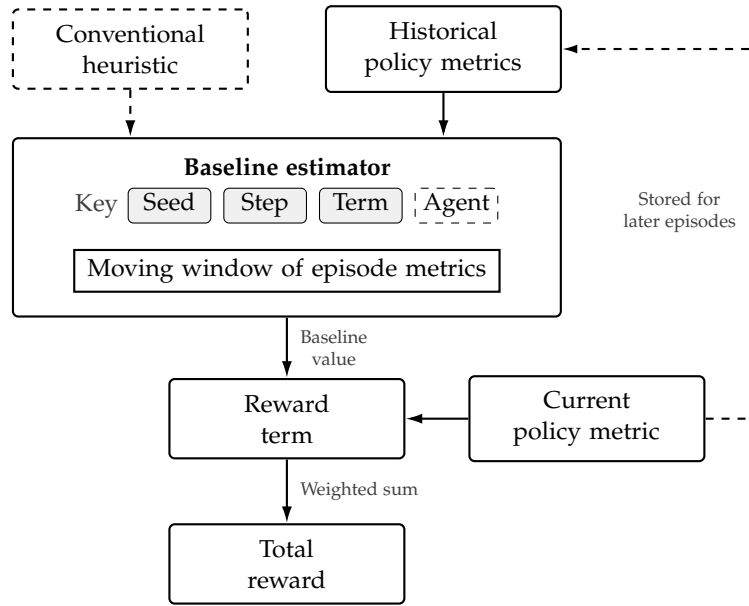


Figure 4.4: The reward-baseline estimator employs a keyed component that is initialized from a conventional heuristic and subsequently updated with metrics from the policy. For each given episode, the current policy metric is normalized by the baseline value to form a single reward term, which contributes to the total reward via a configured weighted sum. The metric is stored after reward computation; therefore, it affects later baseline estimates. The key composition consists of the seed, decision step, and reward term for single-agent and global estimators, while a dashed agent field is included only for local multi-agent estimators.

Listing 4.1: Episodic REINFORCE training loop

```

1 Initialize policy network  $\pi_\theta$  and optimizer
2
3 for each episode  $e$  do
4     Reset environment and get initial observation  $s_0$ 
5     Initialize empty lists  $\mathcal{R}$  and  $\mathcal{L}$ 
6
7     for  $t = 0, \dots, T-1$  do
8          $(\mu_t, \sigma_t) \leftarrow \pi_\theta(s_t)$ 
9         Sample action  $a_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$ 
10        Clip action  $a_t \leftarrow \text{clip}(a_t, 0, 1)$ 
11        Store log-probability  $\ell_t \leftarrow \log \pi_\theta(a_t | s_t)$ 
12        Apply action  $(s_{t+1}, r_t, \text{done}) \leftarrow \text{Env.step}(a_t)$ 
13        Append  $r_t$  to  $\mathcal{R}$  and  $\ell_t$  to  $\mathcal{L}$ 
14
15        if done then
16            break
17        end if
18    end for
19
20    Compute returns  $G_t \leftarrow \sum_{k=t}^{T-1} \gamma^{k-t} r_k$ 
21    Normalize returns  $\hat{G}_t \leftarrow (G_t - \text{mean}(G)) / (\text{std}(G) + \epsilon)$ 
22    Compute loss  $\mathcal{J}(\theta) \leftarrow -\sum_{t=0}^{T-1} \ell_t \hat{G}_t$ 
23    Update parameters  $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{J}(\theta)$ 
24 end for

```

parameter configurations. During trajectory collection, we store the log-probability $\ell_t = \log \pi_\theta(a_t | s_t)$, the environment observation s_{t+1} , and the reward r_t in episode-wise lists $\mathcal{L}$ and $\mathcal{R}$. Upon episode completion, we compute the Monte Carlo return G_t as the discounted sum of future rewards, thereby avoiding bootstrapped value estimates and eliminating critic-induced gradient interference. To stabilize training, we normalize the returns $\hat{G}_t$ episode-wise using a running mean and standard deviation, which reduces gradient variance without altering the optimization direction. The policy loss $\mathcal{J}(\theta)$ is then computed as the negative dot product of log-probabilities and normalized returns, and parameters θ are updated via standard gradient descent. All symbols and their corresponding meanings are summarized in Table 4.2.

Table 4.2: REINFORCE training loop symbols

Symbol	Meaning
π_θ	Stochastic policy network
θ	Trainable policy parameters
e	Episode index
t	Step index inside one episode
T	Episode length
s_t	Observation at step t
s_0	Initial observation after reset
s_{t+1}	Next observation after action
μ_t	Gaussian action mean
σ_t	Gaussian action standard deviation
a_t	Sampled continuous action
ℓ_t	Log-probability of sampled action
r_t	Reward after applying action
$\mathcal{R}$	Episode reward list
$\mathcal{L}$	Episode log-probability list
γ	Return discount factor
G_t	Monte Carlo return from step t
$\hat{G}_t$	Normalized return
ϵ	Numerical stability constant
$\mathcal{J}(\theta)$	Loss
α	Learning rate

Table 4.3: Action space parameter ranges

Symbol	$Q_{\text{Hyst}}^{C_i}$	$Q_{\text{Offset}_{\text{freq}}}^{C_i \rightarrow C_j}$	$Q_{\text{rxlevmin}}^{C_i}$	$T_{\text{XHigh}}^{C_i \rightarrow C_j}$	$T_{\text{XLow}}^{C_i \rightarrow C_j}$	$T_{\text{SLow}}^{C_i}$
Range	$[0, 30]$	$[0, 30]$	$[-100, 0]$	$[-100, 0]$	$[-100, 0]$	$[-100, 0]$

CONTINUOUS ACTION SPACE. For each of the six shared cell (re)selection parameters, an agent outputs a mean μ_i and standard deviation σ_i that define a Gaussian distribution [61, Ch. 13.7]. The full flow of transforming these distribution parameters into a physical action is depicted in Figure 4.5. An agent samples actions from these distributions to adaptively adjust the parameters, resulting in a continuous action vector of size $2 \cdot 6$. The first half of the vector contains the means, while the second half contains the standard deviations. By controlling both the mean and the standard deviation, an agent adaptively modulates its exploration intensity. The sampled actions, constrained to $[0, 1]$, are linearly mapped to the physical parameter ranges (see Table 4.3). We intentionally exclude cell priorities from the action space because they induce discrete, high-impact changes in the reselection logic and are typically adjusted on longer time scales by network operators; thus, priorities are fixed to the reference values of the conventional heuristic baseline.

4.3 OPTIMIZATIONS

To improve training stability, accelerate convergence, and enhance model generalization, we incorporate several key optimization techniques into our RL pipeline.

HISTORY STACKING. The system involves two key time scales:

1. the agent control frequency, referred to as Parameter-Refresh Interval (PRI), which is the period between cell parameter reconfigurations and
2. the UE state dynamics, or UE State Dwell Time (SDT), representing the average number of steps a UE remains in CONNECTED or IDLE mode.

The relationship between these scales is characterized by three cases, based on the ratio of PRI to SDT, referred to as Update-to-Dwell Ratio (UDR). In the case of $\text{UDR} < 1$, rapid adaptation

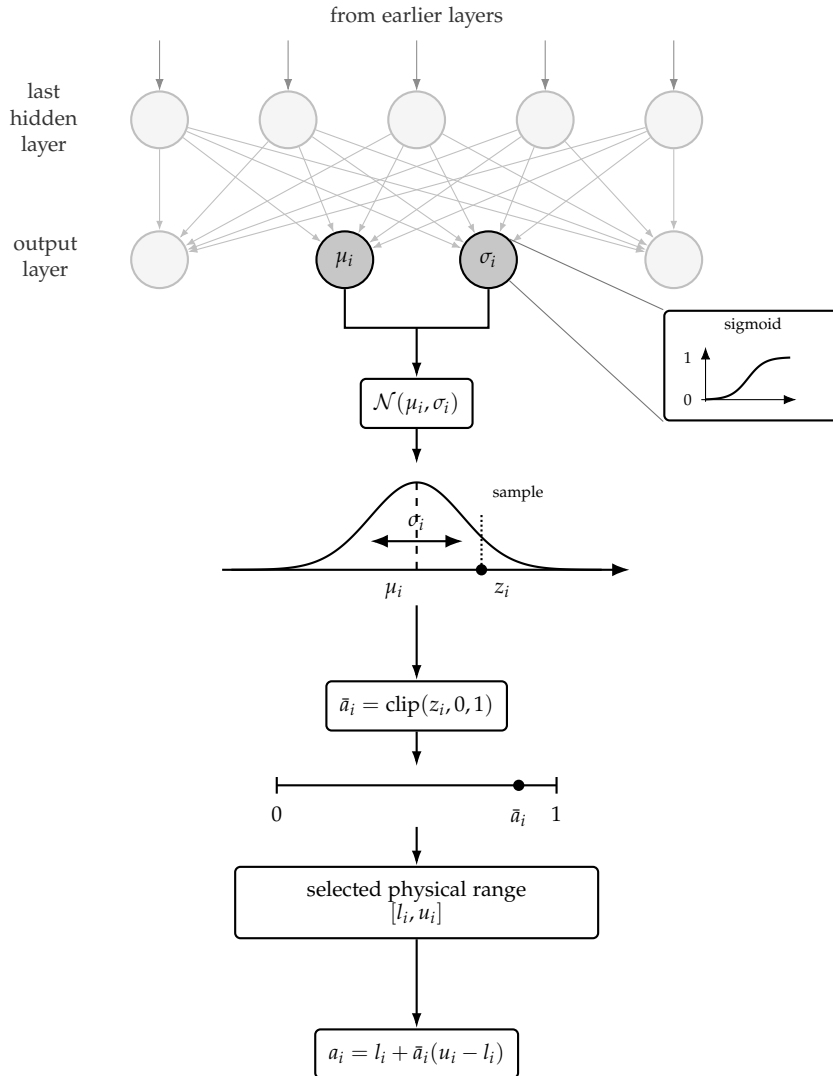


Figure 4.5: Extraction of a single continuous action parameter from the policy output. The highlighted output neurons yield the mean μ_i and standard deviation σ_i of a Gaussian distribution, from which a continuous value z_i is sampled. This sample is clipped to the normalized action interval $[0, 1]$ and then mapped to the corresponding physical parameter range $[l_i, u_i]$.

is enabled via frequent reconfiguration relative to UEs mode changes; for $UDR \approx 1$, dynamics are synchronized due to comparable time scales; and for $UDR > 1$, responsiveness may be limited by slower agent actions. To address partial observability (see Sections 5.1 and 6.1) and manage these temporal dynamics, we augment the observation space with a history of the last h full observations and maintain stable decision-making across varying UDRs. This history stacking allows the agent to disambiguate identical state configurations, such as identical cell-load vectors arising from rising versus falling traffic trends [32, 38, 49].

NEURAL NETWORK ARCHITECTURE. To maintain computational efficiency, we employ a flat, normalized input vector that enables the use of a compact Multi-Layer Perceptron (MLP) shown in Figure 4.6 rather than more complex architectures, like Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM). This design significantly reduces the parameter count. We adopt a simple four-layer feedforward network to ensure fast execution. The input layer aligns with the observation space dimensions, followed by two hidden MLP layers utilizing tanh activation functions as shown in Figure 4.6, enabling it to model non-linear relationships. The output layer applies a sigmoid activation function to constrain the actions within their valid range. Figure 4.6 depicts the neural network architecture employed by the policy.

WARM START. To favor known regions of the action space, we initialize the network weights by configuring the output-layer biases to reproduce the parameter set of the conventional heuristic. This way we build upon their baseline performance and benefit from an established performance baseline that serves as a reliable foundation for subsequent optimization. This warm start strategy helps the policy adapt quickly to the environment and can improve initial training performance [59].

OPTIMIZER. The training procedure applies the AdamW [41] optimizer to update network parameters. By treating weight decay separately from the adaptive learning rate updates, the optimizer strengthens regularization [67] and enhances the generalization properties of the learned policy.

GRADIENT CLIPPING. To stabilize learning and prevent overly large policy updates, we apply global gradient clipping

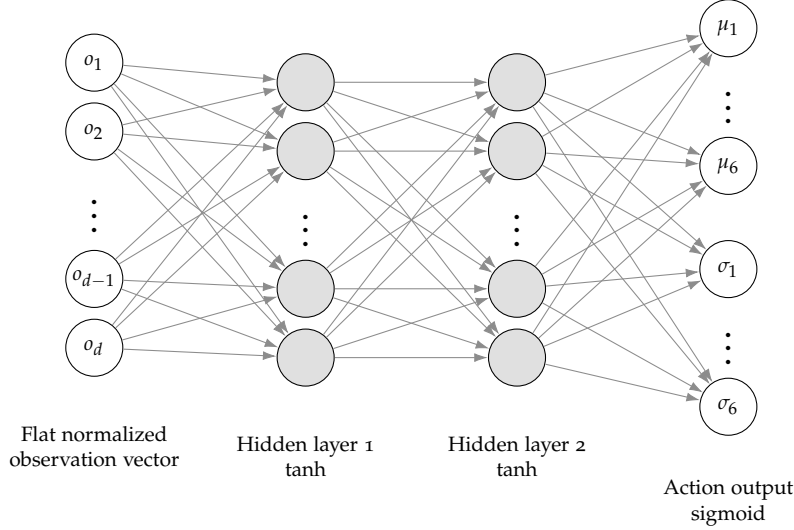


Figure 4.6: Feedforward neural network architecture employed by the policy. Hidden layers utilize tanh activations to capture non-linear transformations, while the output layer applies a sigmoid activation to bound the action outputs to $[0, 1]$. The lightweight design supports efficient inference, enabling real-time execution.

with the ℓ_2 norm [56]. This technique ensures that policy updates remain within a reasonable range, thereby preventing divergence and improving the overall stability of the training process.

SEED SHUFFLING. We shuffle the episode seed order during training to prevent scenario memorization, and we evaluate the agent using unseen seeds to test generalization on unique scenarios. The environment resets after each episode.

CURRICULUM LEARNING. Initially, the untrained policy may take poor actions, leading to sub-optimal resource allocation. This creates cascading effects because UEs that enter CONNECTED mode early can occupy resources under suboptimal allocations, reducing the resources available to UEs that become active later. Such early sub-optimal decisions may push the policy into regions of the state space from which recovery is difficult or impossible. Compounding this problem, the agent is forced to make decisions based on an already degraded state, effectively learning from trajectories that would not occur under an optimal policy and thus further preventing convergence. To mitigate this, we employ curriculum learning [19] by initially training on

shorter, simpler episodes. This approach stabilizes early learning and prevents the agent from accumulating sub-optimal resource allocations. We gradually increase the environmental complexity by extending the episode length over successive rounds to allow the agent to learn from longer episodes. During each episode, we repeat each environment seed three times and shuffle the episodes within each round to expose the agent to a variety of scenarios and prevent overfitting to specific episode sequences. The initial learning rate is set to 10^{-4} and is halved at the start of each new curriculum round.

4.4 GENERALIZATION

Generalizability remains essential for the practical deployment of RL solutions in real-world systems. Generalization challenges must be systematically addressed through a structured set of architectural and training design requirements.

CROSS-SCENARIO GENERALIZATION. To promote cross-scenario generalization, the training pipeline can use domain randomization across a diverse set of unique seeds that feature diverse UE positions and traffic patterns. The randomization of UE locations, activity patterns and mobility profiles generates distinct load patterns in every training episode. Exposure to this broad spectrum of scenarios during training enables the policy to develop robust cross-traffic generalization capabilities rather than memorizing specific deployment configurations.

CROSS-SEED GENERALIZATION. The credit assignment needs to effectively eliminate per-scenario bias (cf. *seed-specific baseline estimators*, Section 4.2). A normalization like this can ensure that gradient magnitudes remain comparable across varying cell topologies and traffic intensities. By focusing the optimization signal on relative improvements rather than absolute metric values, a formulation can facilitate reliable cross-seed transfer and stabilize training across heterogeneous environments with differing spatial configurations and traffic intensities.

GENERALIZATION TO DIFFERENT LOADS. Although training may occur at a fixed reference load, such as 500 UEs, the policy must generalize effectively across varying traffic intensities. A load-invariant formulation can enable seamless transfer

to other load levels—for instance, 100 or 1000 UEs—without requiring any changes to the observation vector dimension or the underlying network architecture.

EPISODE-RELATED GENERALIZATION. Episode-related generalization can be addressed through a curriculum learning strategy that begins with short episodes and gradually extends their duration. This progression stabilizes early learning by allowing the agent to master short-horizon tasks before encountering longer sequences that require planning over slower network phenomena. Consequently, the policy can gain the flexibility to adapt to varied episode lengths and control periods during inference.

To prevent the agent from memorizing the chronological order of training scenarios, the seed sequence can be randomized during each training round. This non-sequential exposure can help accelerate convergence and improve overall generalization by mitigating overfitting to specific episode sequences.

ROBUST POLICY DESIGN. By employing a small model, the risk of overfitting to specific training scenarios can be reduced. This choice can lead to lower variance and markedly less overfitting to rare seeds, as the model has fewer parameters to adjust, which aligns with the bias-variance tradeoff [31, 63].

Idle-mode reselection is a fundamental procedure in mobile network operation, yet it has received limited attention in academic research over more than a decade. Chang et al. [22] extend standard LTE idle-mode cell reselection by incorporating Reference Signal Received Quality (RSRQ)-aware thresholds alongside conventional RSRP measurements, relying on field data to configure these parameters. Malmirchegini et al. [43] formulate multi-carrier layer management as a probabilistic parameter-optimization problem, deriving 3GPP reselection thresholds and offsets from RF propagation statistics to achieve target traffic distributions across carriers. These works represent the sparse literature on the topic, predominantly relying on offline measurements or static optimization rather than adaptive, online decision-making.

The rest of this chapter is structured as follows:

- Section 5.1 formalizes the single-agent control problem by defining the observation and action spaces, formulating a composite reward function that balances throughput, load balancing, and per-UE efficiency, and specifying the checkpointing and convergence criteria used to stabilize training.
- Section 5.2 specifies the experimental configuration and temporal dynamics, evaluates the performance and generalization capabilities across varying network loads, topologies, and control loops, and analyzes the contributions of curriculum learning, training diversity, and mobility to overall policy robustness.
- Section 5.3 summarizes the key findings, discusses the practical implications of transitioning to RL-driven autonomous operations, and outlines directions for future research.

5.1 MODEL

This section formalizes the single-agent control problem by defining the observation and action spaces, the composite re-

Table 5.1: Single-agent observation components

Per-cell i	Global
$(BW_a^i, \rho_{\text{conn}}^i)$	$\mu_a, \sigma_a, \mu_u, \sigma_u, \rho_{\text{idle}}$

ward function, and the training convergence criteria. We detail how the agent perceives network states, selects parameter configurations, and receives feedback to optimize its policy.

OBSERVATION SPACE. The learning problem is formally modeled as a POMDP because the agent lacks full state information, specifically omitting variables like per-UE mobility, IDLE/CONNECTED timers, and reselection counters. The observation space comprises per-cell features and global metrics, summarized in Table 5.1. Per-cell features consist of the normalized available bandwidth BW_i^a and the connected ratio ρ_i^{conn} , defined as the number of connected UEs divided by the total number of UEs, including the UEs camping on it in IDLE mode. Global metrics aggregate these features into a fixed-size representation by computing the mean and standard deviation of the normalized available bandwidth, μ_a and σ_a , and the mean and standard deviation of the normalized connected UEs count, μ_u and σ_u , alongside the idle ratio ρ_{idle} . This fixed-size representation compresses the per-cell data into a bird’s-eye view of the network state and enables policy generalization across varying scales.

ACTION SPACE. CellPilot assumes homogeneous cell parameters across the network region, reflecting current operator practice due to the complexity of manual configuration. Extending the framework to heterogeneous parameters remains feasible through expanded state-action spaces or a multi-agent approach, which is covered in Section 6.1.

REWARD. The reward function is designed to optimize three competing objectives: total network throughput, load balancing, and per-UE efficiency. We formulate the composite reward as

$$r_t = w_1 \cdot r_t^{\text{TP}} + w_2 \cdot r_t^{\text{LB}} + w_3 \cdot r_t^{\text{UE_EFF}}, \quad (5.1)$$

where the weights sum to unity. The throughput component r_t^{TP} rewards improvements in total network throughput across all cells. The load-balancing component, r_t^{LB} , encourages equalizing

throughputs across cells by penalizing the standard deviation σ_t of per-cell throughputs relative to a baseline $B^\sigma(t)$. This aligns with the operator objective of distributing network load evenly to prevent both cell congestion and resource underutilization. The per-UE efficiency component $r_t^{\text{UE_EFF}}$ maximizes the mean per-UE throughput while scaling the reward by the fraction of active UEs. This design prevents excessive rewards for trivial scenarios with few active UEs and explicitly rewards performance gains in more challenging high-load scenarios.

Consequently, the baseline-normalized reward terms are formulated as follows: The throughput term is defined by

$$r_t^{\text{TP}} = \left(\frac{T_t}{B^{\text{TP}}(t)} \right) - 1 \quad (5.2)$$

where T_t denotes the accumulated total throughput across all cells at time step t , and $B^{\text{TP}}(t)$ is a moving average baseline estimator. To capture the gain in load balancing, we define the reward component as

$$r_t^{\text{LB}} = \left(\frac{B^\sigma(t)}{\sigma_t} \right) - 1. \quad (5.3)$$

Here, σ_t measures the standard deviation of per-cell throughputs, while $B^\sigma(t)$ serves as the reference baseline estimator. The baseline ratio is inverted for the load-balancing term to maintain a consistent reward sign convention, ensuring that positive values consistently denote improvement despite load balancing requiring a reduction in throughput variance rather than an increase. Finally, the per-UE efficiency term is given by

$$r_t^{\text{UE_EFF}} = \left(\frac{\bar{T}_{\text{UE},t}}{B^{\text{UE}}(t)} - 1 \right) \cdot \min \left(1, \frac{\bar{\text{UE}}_{\text{active},t}}{\text{UE}_{\text{max}}} \right) \quad (5.4)$$

where $\bar{T}_{\text{UE},t}$ is the mean per-UE throughput, $B^{\text{UE}}(t)$ is its baseline estimator, and $\bar{\text{UE}}_{\text{active},t}$ is the average number of active UEs during interval t . The term UE_{max} denotes the total number of UEs in the network, and the min function scales the reward to account for current load. This design encourages the agent to maximize throughput in high-load network conditions with more active UEs, preventing it from focusing on a few UEs at the expense of others by ensuring the reward signal scales proportionally with the active population size.

CHECKPOINTING AND CONVERGENCE. Naturally, agent performance fluctuates throughout the training process. To ensure

that the optimal policy is preserved, we implement a checkpointing mechanism that periodically saves the current model. Training convergence is determined when episode rewards approach zero, indicating that the agent no longer improves upon its historical performance.

5.2 EVALUATION

This section evaluates the performance of CellPilot in real-world mobile network scenarios, focusing on three key objectives:

- quantifying the performance gains of CellPilot compared to the conventional heuristic approach;
- assessing its ability to generalize to new, unseen network scenarios; and
- identifying the specific contribution of major optimizations within CellPilot to these gains.

5.2.1 Evaluation Setup

The learning mechanisms follow the common framework established in Chapter 4; here, we specify the concrete parameter instantiations and configurations required for reproducible training.

We train the policy over 100 distinct episodes, where each episode employs a unique random seed to determine UE positions, mode transitions, mobility patterns, and other stochastic environmental aspects. Evaluation is conducted on 100 unseen seeds for each scenario to assess generalization performance across diverse random realizations. The UE mode switch is modeled as a Poisson process with exponential interarrival times parameterized with rates $\lambda_{\text{IDLE}} = \lambda_{\text{ACTIVE}} = 0.2$, leading to an average SDT of 5 steps. The observation vector includes history stacking of the last $h = 10$ observations. Action space ranges are normalized according to Table 3.2. Exploration is managed using a Gaussian distribution with the standard deviation σ capped at 0.1 (see Figure 4.5). The standard reward weights are set to $w_1 = 0.4$ for throughput, $w_2 = 0.4$ for load balancing, and $w_3 = 0.2$ for per-UE efficiency. For the scalability scenario using the large geographical area and synchronous update scenario evaluating temporal dynamics, we employ weights of $w_1 = 0.025$, $w_2 = 0.95$, and $w_3 = 0.025$. Because the agent

achieved disproportionately large throughput gains in these scenarios, we increased the load-balancing weight to mitigate this imbalance and incentivize progress across all reward components. Each reward component maintains a separate baseline estimator. The models employ 1024 neurons per layer with 2 hidden layers, as using 512 neurons per layer led to unstable training and suboptimal performance. This lightweight architecture requires fewer than 5 MFLOPs per inference pass, enabling sub-millisecond execution on commodity hardware. We apply a gradient clipping threshold of 10 to ensure training stability. The initial learning rate is set to 10^{-4} and is halved at each new curriculum-learning round, while the episode length starts at 30 steps and increases by 5 steps per round across 4 total rounds, resulting in a final episode length of 45 steps.

TEMPORAL DYNAMICS. We evaluate the RL agent under different temporal dynamics to assess its robustness and adaptability. The baseline model is trained and evaluated with a PRI of 1, resulting in a reference UDR of approximately 0.2. In the *Stress Test* scenario, the model is evaluated using a PRI of 10 and a UDR of approximately 2 to test robustness against unseen dynamics, where an average of two UE mode changes occur between updates. The *Synchronous Updates* scenario is trained with a PRI of 5 to match the expected SDT, resulting in a UDR of approximately 1. Finally, the *Slow Updates* scenario employs a PRI of 10 and a UDR of approximately 2 to simulate a slow control loop. The key distinction from the *Stress Test* is that, unlike the latter which tests unencountered dynamics, the *Slow Updates* scenario involves both training and evaluation using these identical temporal parameters.

5.2.2 Overall Performance and Generalizability

We evaluate the performance of CellPilot across three dimensions:

- across three geographical areas
- under load scaling from 100 to 1000 UEs, and
- over four distinct temporal dynamics ranging from low-latency to slow control loops (comprising *Baseline*, *Stress Test*, *Synchronous Updates*, and *Slow Updates*)

These evaluations probe the spatial, load-dependent, and temporal generalization capabilities of the agent. Each evaluation utilizes 100 unseen scenarios with 500 UEs and a fixed episode length of 60 steps. For load scaling, we evaluate the baseline area with $N = 100, \dots, 1000$ UEs in increments of 100. The performance is compared against the conventional approach using the reference values provided in Table 3.1. For easier visual interpretation, load-balancing results are reported using an improvement-oriented sign convention, so reductions in per-cell throughput standard deviation appear as positive gains.

CONVERGENCE. Because the reward signal is normalized against seed-specific baseline estimators, the agent receives positive feedback only when its performance exceeds historical expectations. Consequently, maintaining static performance yields diminishing returns as the moving-average baseline progressively converges toward the agent’s recent outputs, requiring continuous improvement to sustain positive reward signals. Naturally, the agent’s reward declines over the course of the training, with convergence defined as the point where the reward approaches zero. Convergence assessment sometimes remains ambiguous, as temporary zero-reward intervals are occasionally followed by continued optimization. To resolve this, we determine convergence by monitoring throughput and load-balancing performance across different model checkpoints.

Additionally, we observe that the agent occasionally experiences temporary stagnation at zero reward before resuming optimization; however, it subsequently focuses exclusively on a single performance metric. We observe that throughput gains are comparatively easier to achieve and larger in magnitude. Therefore, we adapt reward weights in specific scenarios to maintain balanced optimization. Figure 5.1 shows the episode reward over training steps for the baseline case. Vertical lines demarcate curriculum-learning rounds where the episode length increases. The reward trace is smoothed using an exponentially weighted moving average with $\alpha = 0.2$. During the initial curriculum round, the normalized rewards fluctuate between 0.02 and -0.004 . Training stabilizes following the second curriculum round, signaling convergence. The standard deviation over the final 100 episodes is 7×10^{-4} .

LOAD SCALING. We quantify performance variation relative to network load by fixing the agent trained on 500 UEs and

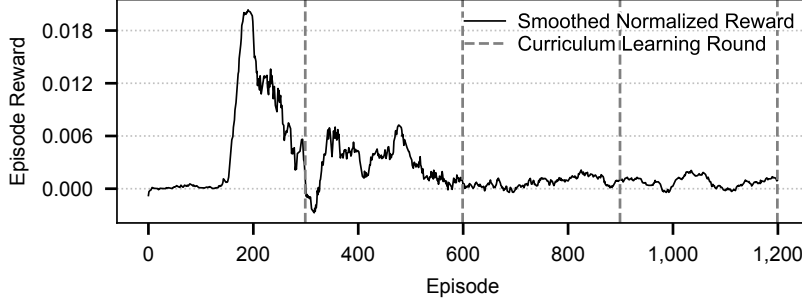


Figure 5.1: Training convergence of the single-agent policy in the baseline area. The figure depicts the exponentially smoothed ($\alpha = 0.2$) normalized episode reward across training episodes, while vertical markers denote curriculum-learning rounds with progressively increasing episode lengths. Adapted from [33], © 2026 IFIP.

evaluating it across $N = 100, \dots, 1000$ UEs in the baseline area. Figure 5.2 presents the reward components as averages over the UE count. Despite being trained exclusively on 500 UEs, the agent generalizes effectively across all evaluated loads, demonstrating consistent gains across every reward component. It consistently outperforms the conventional heuristic in throughput, load balancing, and per-UE efficiency. The observed gains are lower for smaller N . Specifically, load balancing improves by up to 26 % at $N = 1000$, total throughput rises consistently with increasing UE count reaching a maximum gain of 118 %, and per-UE efficiency improves by up to 59 % at $N = 1000$. Higher UE counts lead to diminishing returns across all components. This trend is expected for throughput and per-UE efficiency because the conventional heuristic increasingly saturates the available spectrum as the UE count grows. Similarly, load balancing improves naturally with more UEs per cell, which reduces the remaining optimization potential for the agent. This occurs because higher UE density per cell reduces relative load fluctuations: as individual connection events average out, the system naturally equalizes cell loads. Because the conventional heuristic already performs well under these conditions, the RL agent has less remaining room to improve the metric. Overall, these results confirm that the learned policy scales effectively with network load, achieving the largest relative gains under high-density conditions.

CROSS-AREA EVALUATION. We evaluate the agent trained on the baseline area on both the baseline and alternative area,

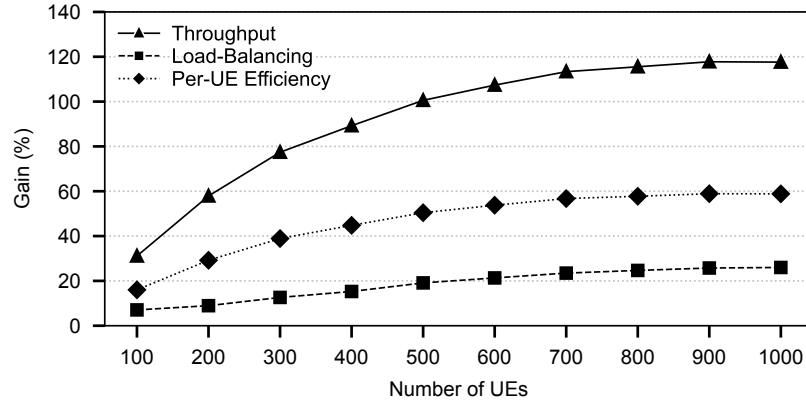


Figure 5.2: Single-agent average performance gains relative to the conventional heuristic in the baseline area under varying UE counts. The policy is trained on 500 UEs and evaluated from 100 to 1000 UEs, demonstrating improved performance across all tested metrics. Adapted from [33], © 2026 IFIP.

which feature different cell topologies and UE positions while sharing geographic size, tower count, and cell count. Figure 5.3 depicts the results of the cross-area evaluation. The agent outperforms the conventional heuristic in both areas across all three metrics. The performance gains in the alternative area remain comparable to those in the baseline area. Smaller confidence intervals for throughput and per-UE efficiency in the alternative area indicate more consistent performance. Occasional negative deviations from the heuristic baseline remain limited in magnitude and frequency, as indicated by the quartile statistics in Figure 5.3. The agent is capable of correcting these deviations in subsequent action steps, thereby recovering performance over time. Because the current formulation considers only idle-mode cell reselection in isolation, these underperformances could be further mitigated by connected-mode handovers, which provide additional optimization opportunities in practice. These results demonstrate that the learned policy generalizes effectively to unseen areas with distinct topologies and UE distributions. We attribute this consistency to the geographical layout of the alternative area, where cells are more evenly distributed compared to the clustered baseline cells, resulting in less diverse UE-to-cell associations. In total, the policy maintains strong robustness to topology shifts.

SCALABILITY TO LARGER AREAS. We evaluate scalability of the formulation by training on the large area that is $2.6\times$ the size of the baseline with $3.2\times$ the cells using $N = 500$ UEs. For

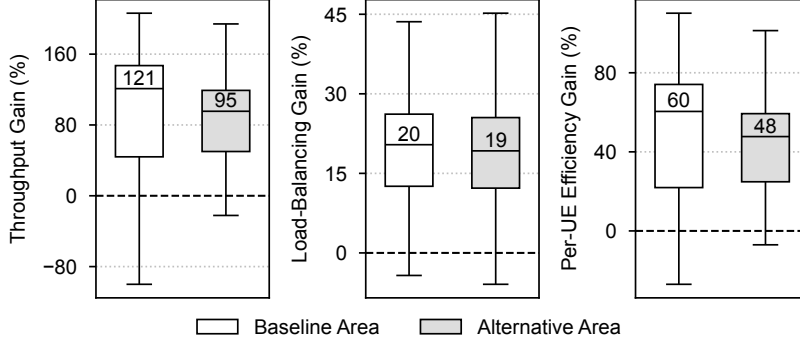


Figure 5.3: Single-agent performance gains relative to the conventional heuristic in the baseline and alternative areas. The boxplots compare throughput, load-balancing, and per-UE efficiency gains across 100 unseen evaluation scenarios, indicating that the policy trained on the baseline area generalizes to a geographically distinct topology. Adapted from [33], © 2026 IFIP.

this scenario, we increase the moving-average estimator window size from 2 to 10 to stabilize the reward baseline. We observed that using a smaller window causes the baseline estimator to adapt very quickly to poor performance, which suppresses the reward signal and leaves the agent with an insufficient learning signal. In contrast, a larger window maintains a higher, slower-adapting baseline, which requires the agent to achieve larger improvements to generate positive rewards or progress. Figure 5.4 depicts the results of the evaluation on the large area. The agent outperforms the conventional heuristic across all metrics with high confidence. The median total throughput gain is 82 %, which is lower than the 121 % observed in the baseline area. This reduction is expected because the heuristic utilizes most of the available spectrum when there are fewer UEs per cell, leaving limited optimization headroom for the agent. Load-balancing improvement shows a median gain of 3 %, approximately 1/6 of the gain seen in the baseline area. This is expected due to the use of homogeneous (re)selection parameters across all cells, which limits fine-tuning capabilities of the agent in a larger area with more cells. The per-UE efficiency gain is 41 % at the median, roughly 2/3 of the baseline gain. These results show that the formulation reliably improves upon the heuristic even in larger topologies, proving its scalability.

DIVERSE TEMPORAL DYNAMICS. Figure 5.5 presents the results under different temporal dynamics. In the *Stress Test*, we

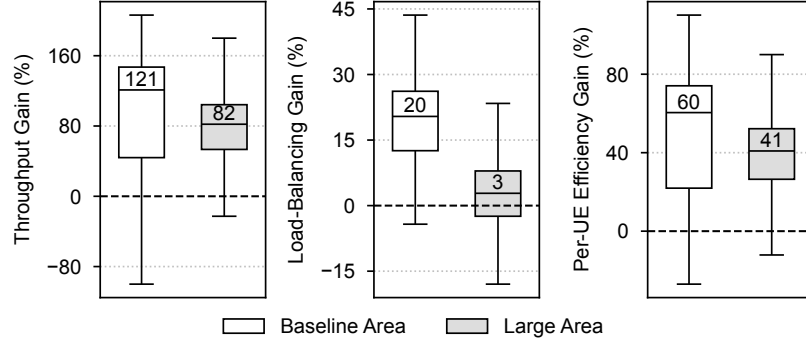


Figure 5.4: Single-agent performance gains over the conventional heuristic in the baseline and large areas. The boxplots compare throughput, load-balancing, and per-UE efficiency gains across 100 unseen scenarios, demonstrating that the learning formulation maintains performance when scaling to a larger topology with additional towers and cells. Adapted from [33], © 2026 IFIP.

observe median gains of 56 % for throughput, 7 % for load balancing, and 27 % for per-UE efficiency. These gains are lower than those in the baseline scenario where the agent is evaluated under the same temporal dynamics as it was trained on. This reflects the temporal mismatch: the agent learned to operate on shorter timescales and now operates under a slower control loop, limiting its ability to react quickly to traffic bursts and update network parameters. Nevertheless, the agent still improves performance across all metrics compared to the conventional heuristic, indicating that it has learned a policy that is robust to changes in temporal dynamics. In the *Slow Updates* scenario, the policy achieves median gains of 139 % for throughput, 20 % for load balancing, and 69 % for per-UE efficiency. This suggests that training on a slower timescale enables the agent to identify more resilient, long-term parameter configurations that generalize better across multiple UE state transitions. The *Synchronous Updates* scenario performs best, with median gains of 147 % for throughput, 21 % for load balancing, and 72 % for per-UE efficiency. These results yield the strongest median performance among all tested temporal dynamics scenarios. This indicates that aligning the update interval with the UE transition timescale allows for optimal parameter update timing, minimizing reactive lag and enhancing performance. However, this also introduces timing sensitivity, as evidenced by the higher variance in throughput gains: if many UEs transition shortly after an update, the new configuration may become stale before the next

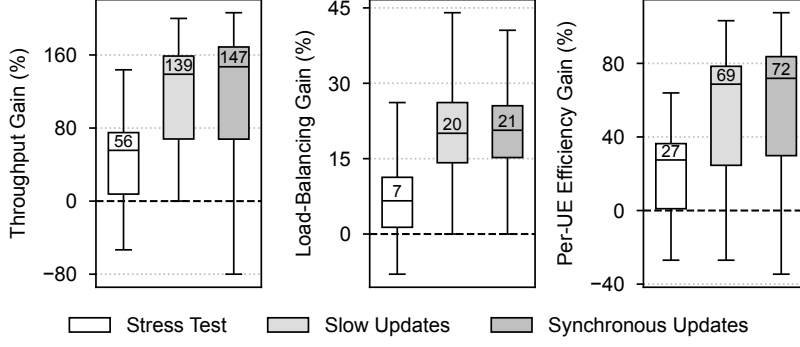


Figure 5.5: Single-agent performance gains relative to the conventional heuristic under different temporal dynamics. The boxplots compare throughput, load-balancing, and per-UE efficiency gains across the stress-test, slow-update, and synchronous-update scenarios, quantifying the effect of different control-loop latencies. Adapted from [33], © 2026 IFIP.

update. Slower update intervals mitigate this effect by acting as temporal regularization, trading short-term responsiveness for stability through averaging over multiple state changes.

CURRICULUM LEARNING. We evaluate the impact of curriculum learning by disabling it and fixing the episode length at 50 steps, corresponding to the final curriculum length. We retain the repeating mechanism, ensuring the agent trains on each unique episode three times in a random order. Figure 5.6 depicts the resulting performance gains across all metrics. The standard deviation of the episode reward over the last 100 episodes is $4.9\times$ higher compared to the curriculum-enabled setup, indicating a substantially slower policy convergence rate without progressive difficulty scaling.

TRAINING DIVERSITY. We evaluate the impact of training seed count by comparing the default 100 seeds against an increased set of 500 seeds. The 500-seed configuration causes noticeably slower convergence due to the expanded scenario diversity, though Figure 5.7 confirms that learning progress remains visible even during the final two curriculum rounds. We assess performance at two checkpoints: an early stage at 2300 episodes with near-zero reward, and the late stage at the end of training. The late checkpoint substantially outperforms the early checkpoint across all three metrics, demonstrating a shift from selective optimization to balanced performance gains. The early 500-seed model exhibits intermediate performance,

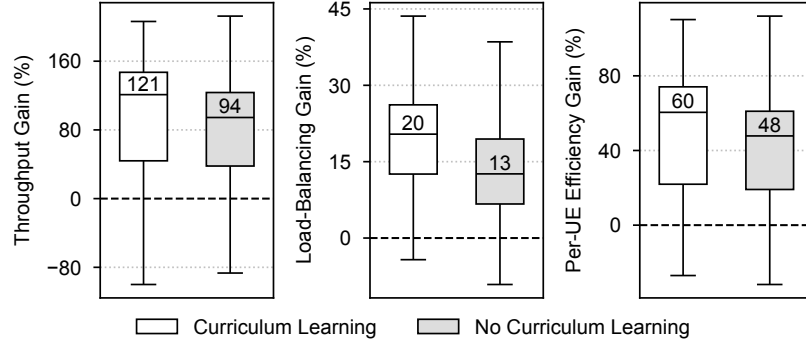


Figure 5.6: Single-agent performance gains relative to the conventional heuristic, with and without curriculum learning. The box-plots compare throughput, load-balancing, and per-UE efficiency gains, indicating that excluding curriculum learning reduces median performance and slows convergence. Adapted from [34].

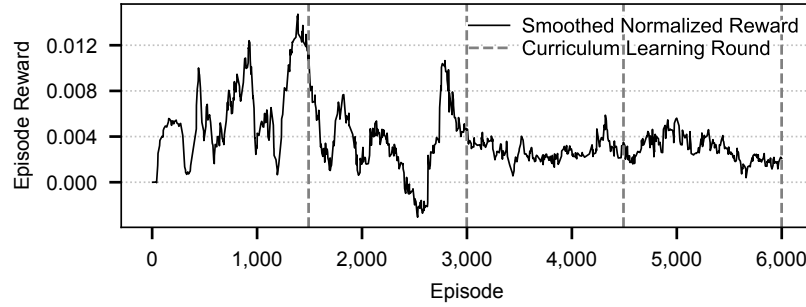


Figure 5.7: Training convergence of the single-agent policy averaged over 500 training seeds. The curve shows the exponentially smoothed ($\alpha = 0.2$) normalized episode reward across training episodes, while vertical markers indicate curriculum-learning rounds with progressively increasing episode lengths. Adapted from [34].

surpassing the 100-seed baseline but falling short of the late 500-seed model. The final checkpoint achieves median gains of 167% in throughput, 39% in load balancing, and 83% in per-UE efficiency. Performance improves consistently across all metrics, with load balancing exhibiting the largest relative gain as expected when training on a more diverse dataset.

UE MOBILITY. We evaluate the impact of UE mobility by enabling movement exclusively during the evaluation phase while retaining static UEs during training. The UEs travel at 30 km/h along street paths. We assess two model variants:

- a baseline model trained on 100 seeds, and

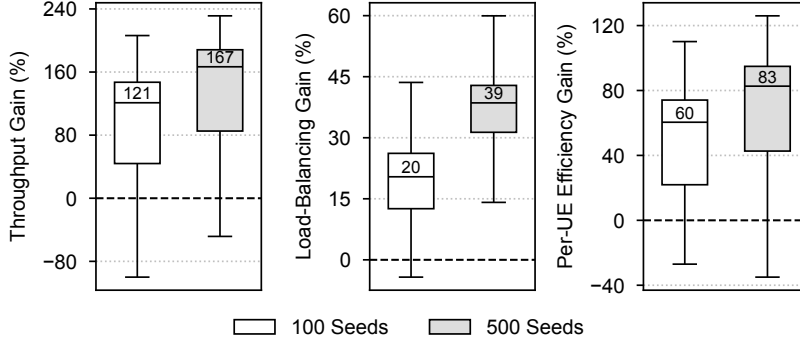


Figure 5.8: Single-agent performance gains relative to the conventional heuristic, trained with 100 and 500 seeds. The boxplots compare throughput, load-balancing, and per-UE efficiency gains, indicating that greater training diversity improves median performance across the evaluated metrics. Adapted from [34].

- a model trained on 500 seeds, allowing us to assess whether increased training diversity improves robustness to dynamic mobility patterns.

Figure 5.9 shows the evaluation results for the baseline model under UE mobility. Despite being trained exclusively on static UEs, the agent still outperforms the conventional heuristic across all metrics, achieving gains of 106 % in throughput, 22 % in load balancing, and 53 % in per-UE efficiency. The consistency of these gains indicates that the agent has learned the underlying environmental dynamics rather than memorizing fixed UE positions, thereby enabling effective adaptation to dynamic UE mobility.

Figure 5.10 shows the results for the 500-seed model evaluated under mobile conditions. Performance further improves across all metrics compared to both the baseline and 100-seed mobility models. The throughput gain increases to 125 %, load balancing rises to 39 %, and per-UE efficiency reaches 62 %. These substantial improvements, particularly in load balancing, indicate that training on a more diverse set of scenarios enhances the agent’s generalization capability to dynamic network conditions such as UE mobility.

5.3 CONCLUSION

This chapter presented CellPilot, a framework for autonomous idle-mode cell (re)selection parameter adaptation in heterogeneous cellular networks. The problem is formulated as a POMDP

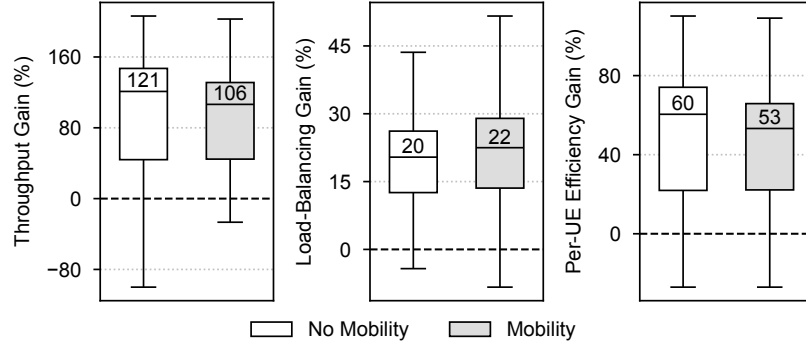


Figure 5.9: Single-agent performance relative to the conventional heuristic, with and without UE mobility during evaluation. The policy is trained on static UEs across 100 seeds and evaluated with UEs moving at 30 km/h along street paths, demonstrating improved performance in throughput, load balancing, and per-UE efficiency. Adapted from [34].

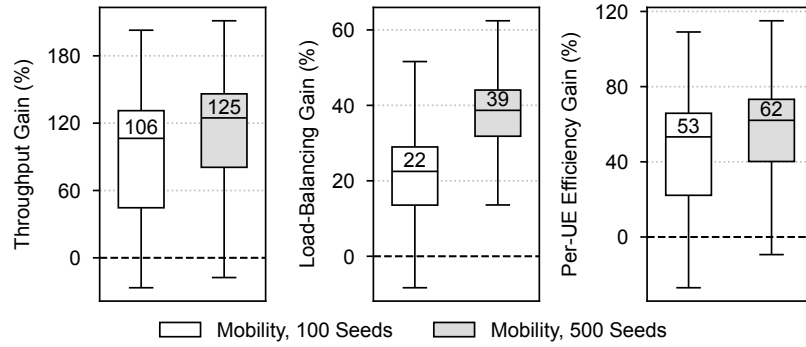


Figure 5.10: Single-agent performance gains relative to the conventional heuristic under UE mobility, for policies trained with 100 and 500 seeds. The boxplots compare throughput, load-balancing, and per-UE efficiency gains, indicating that increased training diversity improves robustness under mobile UE conditions. Adapted from [34].

and addressed using optimization techniques that incorporate scale-invariant observation spaces, REINFORCE with seed-specific baselines, curriculum learning, and history stacking. This formulation enables the agent to optimize network performance under highly dynamic conditions. Evaluation via a custom simulator deployed on real-world topologies demonstrates that CellPilot outperforms conventional heuristics by up to 167 % in throughput. Performance remains robust across unseen geographic regions, varying user loads from 100 to 1000 UEs, and heterogeneous control loop latencies. This strong generalizability is achieved without requiring any retraining phase. The design scales effectively to larger deployments, as validated with a 48-cell topology. Because it operates strictly through standard parameter reconfiguration, CellPilot remains fully compatible with existing 4G/5G specifications. By establishing a reliable mechanism for adaptive cell (re)selection, this work demonstrates the practical feasibility of transitioning from static configurations to RL-driven autonomous operations, thereby laying the groundwork for self-optimizing RAN.

COOPERATIVE MULTI-AGENT RL FOR IDLE-MODE CELL (RE)SELECTION

Idle-mode cell (re)selection governs how **UEs** determine which cells to camp on while in IDLE mode. This mechanism dictates both cell suitability evaluation and reselection timing, thereby directly influencing user experience and overall network performance. Although 3GPP standards formally define the underlying decision logic [7, 8], they delegate the configuration of key thresholds and control parameters to network operators. Consequently, a practical optimization gap emerges, as the standards provide no explicit guidance for selecting parameter values that effectively match specific deployment characteristics. Operators typically resolve this gap through manual engineering and iterative tuning, producing conventional heuristic parameter sets that we adopt as the operational reference baseline throughout this chapter (see Section 3.2).

Prior work (see Chapter 5) demonstrates the feasibility of applying **RL** to idle-mode cell (re)selection through a centralized single-agent formulation with homogeneous area-wide parameterization [33]. This approach employs a single shared parameter vector across all cells, with a centralized controller making decisions based on global observations. While this centralized design achieves strong gains over the operator-provided heuristic baseline, it falls short of the envisioned practical deployment architecture.

Real deployments exhibit cell-level heterogeneity across load, coverage, radio conditions, and **UE** distributions. This heterogeneity motivates cell-specific parameter control, as a single area-wide parameter vector imposes a global compromise that may be unsuitable for individual cells. Because idle-mode reselection is inherently coupled, suboptimal parameterization in one cell propagates system-wide: an overloaded cell pushes **UE** traffic to adjacent cells, which may become congested or subsequently overloaded. The resulting load distribution therefore emerges from the combined parameter settings of neighboring cells rather than from any single cell's configuration. This coupled effect amplifies the limitations of homogeneous parameterization, since a uniform vector must impose a global com-

promise that inadequately addresses heterogeneous local conditions. Consequently, practical operation requires distributed, cell-specific control that adapts parameters to local heterogeneity while accounting for cross-cell load coupling.

However, heterogeneity alone is insufficient to make the centralized formulation practical, because a centralized controller would still require global state at decision time. It must also output a joint action over all cell-specific parameter vectors. With each additional cell, the joint action space expands proportionally, making exploration computationally expensive, complicating credit assignment, and restricting deployment scalability.

The proposed architecture therefore adopts a distributed multi-agent mechanism, assigning one decision-making agent to each cell. Each agent outputs its own cell-specific parameter vector based on local and neighborhood observations, which keeps the per-agent action dimension fixed and better matches scalable cell-level operation. Inter-cell coordination is achieved through neighborhood-aware observations, shared coordination terms in the reward, and training-stabilization mechanisms. This design targets a more deployable control architecture while preserving the main performance benefits observed with centralized RL-based reselection control.

This shift improves operational fit but alters the learning problem structure, as each agent acts from partial local and neighborhood observations and must achieve coordination without centralized joint control. Global information sharing is therefore limited to avoid excessive synchronization overhead. We address these challenges by giving each agent a fixed-size observation over its own cell and configured neighbors, combining local rewards with compact aggregate feedback for network-wide objectives, and stabilizing training through reward-baseline preservation, warm start, history stacking, and gradient clipping. Table 6.1 outlines the key differences between the centralized, area-wide formulation and the distributed, cell-specific design.

The single-agent formulation presented in Section 5.1 establishes the foundational RL framework for autonomous idle-mode parameter adaptation. This chapter introduces a multi-agent formulation that expands the optimization perspective from network-level aggregates to explicitly include UE throughput. By incorporating a neighborhood-aware reward, the design explicitly optimizes UE QoE in CONNECTED mode, ensuring that idle-mode reselection decisions directly translate into measurable per-user performance gains.

Table 6.1: Comparison of single- and multi-agent formulation

Property	Prior Formulation	Proposed Formulation
Control	Centralized	Distributed
Parameters	Homogeneous	Heterogeneous
Observations	Global	Local + neighborhood
Action output	Area-wide	Cell-specific
Reward structure	Global objective	Distributed coordination
Scalability	Limited	Scalable

The rest of this chapter is structured as follows:

- Section 6.1 defines the distributed cell-agent formulation, including the neighborhood setup, message exchange assumptions, neighborhood-aware observation space, and cell-specific action space.
- Section 6.2 details the composite reward structure that combines local cell objectives, neighborhood-level UE QoE, and lightweight global coordination terms for throughput and load balancing. It also introduces permanent baseline anchoring and checkpoint selection for stable training.
- Section 6.3 shows that the proposed formulation consistently improves over the conventional heuristic, remains effective under neighborhood-only information, UE mobility, slower control, changed topology, different load levels, and increased training diversity, and retains performance comparable to the prior centralized formulation under the same slow control loop. It further analyzes UE QoE improvements using throughput-distribution Empirical Cumulative Distribution Functions (ECDFs).

6.1 MODEL

The adoption of a distributed multi-agent control architecture introduces inherent challenges to learning stability and inter-agent coordination. Each cell agent operates under a restricted observation scope, receiving only a local network view rather than global state information. Because coordination must emerge without an omniscient controller, the overall design complexity increases. Accordingly, both the state and reward signals

remain partially observable and locally restricted per agent. This architectural constraint forces each agent to learn a distinct policy that selects cell-specific parameter values. While such parameter heterogeneity enables effective local adaptation, it simultaneously complicates joint behavior coordination and training stabilization.

This section defines the cooperative multi-agent model used for distributed idle-mode cell (re)selection. It specifies the cell-agent assignment, neighborhood construction, message-exchange assumptions, neighborhood-aware observation space, and cell-specific action space. The reward design and training-stabilization mechanisms are then covered in Section 6.2.

NEIGHBORHOOD-AWARE MULTI-AGENT FORMULATION.

We adopt a neighborhood-aware multi-agent formulation in which exactly one agent is assigned to each cell to perform cell-local parameter updates. Each agent operates on local information comprising its own cell state and a small neighborhood of related cells, which restricts inter-cell coordination to plausible local communication rather than full global message exchange. This design aligns with practical cellular operation, where base stations exchange limited neighbor-related information and do not maintain full global state at decision time. Both the observation and reward structures are designed to be neighborhood-aware: each agent observes self-cell metrics, neighbor-cell metrics, and a compact shared summary of the wider network state, while the reward combines local control objectives with a neighborhood-level UE quality-of-experience term and shared global coordination terms. To validate that the limited shared summary provides measurable benefits rather than serving as the primary design element, we evaluate a local-information-only variant in Section 6.3.2.

NEIGHBORHOOD SETUP. A neighborhood is defined as a set of cells that could plausibly serve the same UEs, meaning their coverage regions overlap in space. In practice, network operators typically possess valid or reasonable neighbor relations, which are precomputed before training and evaluation. Within the simulator, these neighbor relations are derived heuristically from propagation and coverage overlap. This derivation may include co-located sectors on the same cell tower or cells from different towers, provided they cover a common area and can therefore compete for the same UEs population. Two cells are classified as

Table 6.2: Multi-agent observation components

Local	Neighborhood	Global
a_i, u_i	(a_j, u_j) for up to 3 strongest neighbors	$\mu_a, \sigma_a, \mu_u, \sigma_u, \rho_{\text{idle}}$

neighbors if they mutually cover each other above a minimum signal-strength threshold and satisfy a beam/coverage check.

MESSAGE EXCHANGE. In 5G/O-RAN architecture, the E2 interface supports near-real-time reporting and control between E2 nodes and the RAN Intelligent Controller (RIC), encompassing cell-level measurements, load-related KPIs, and control commands [23]. Similarly, in 4G/LTE architecture, X2AP provides load-management procedures for exchanging load, overload, and resource-status information between neighboring Evolved Node Bs (eNBs) [4]. Our formulation requires only a fixed-size exchange per decision round, consisting of local cell summaries, up to three neighbor summaries, and a few global aggregate indicators. This design avoids transmitting per-UE measurements or full network state. Consequently, the communication overhead remains small and aligned with existing neighborhood-level coordination mechanisms.

OBSERVATION SPACE. Each cell-agent i observes a combination of information from its own cell, neighbor-cell metrics, and global aggregated indicators. For its own cell i , the agent observes the available-bandwidth ratio a_i and the connected-UE ratio u_i . For up to $k = 3$ strongest-signal neighbor cells, the agent observes the available-bandwidth ratio a_j and connected-UE ratio u_j per neighbor j . For global aggregated metrics, the agent observes the mean μ_a and standard deviation σ_a of the available-bandwidth ratio across all cells. It also observes the mean μ_u and standard deviation σ_u of the connected-UE ratio across all cells, along with the idle-UE ratio ρ_{idle} . One snapshot of this observation has a size of ≤ 13 , as detailed in Table 6.2. To provide temporal context, the agent stacks the last 10 observation snapshots (see *history stacking* in Section 4.3). This stacking results in a full per-agent observation size of 130.

ACTION SPACE. The action space is continuous, where each agent controls six cell (re)selection parameters for its cell. For each parameter, the model outputs a mean μ_i and standard

deviation σ_i . The action is sampled from Gaussian distributions and mapped to the corresponding physical parameter ranges, which are listed in Table 3.2. Cell priorities are fixed to the heuristic baseline and remain constant.

6.2 LEARNING

This section presents the policy learning framework and stabilization techniques required for the cooperative multi-agent formulation. The training procedure integrates an RL algorithm with a reward structure that balances local cell performance and global network coordination through a composite, baseline-normalized metric. A policy checkpointing mechanism is introduced to preserve optimal trade-offs across competing objectives and enable reproducible model selection. Policy architecture and training configurations are explicitly designed to promote robustness and generalization across heterogeneous deployment scenarios.

REWARD. To balance local and global objectives, the reward is delivered at every decision step based on the agent's previous action and the resulting network state. This reward structure comprises two local components that depend on the agent's own cell and its neighborhood, alongside two shared global components that reflect overall network performance. Consequently, the reward functions as a composite metric with four distinct objectives. The per-agent reward for agent i is defined as

$$r_i = w_1 \cdot r_i^{\text{Cell_TP}} + w_2 \cdot r_i^{\text{UE_QoE}} + w_3 \cdot r^{\text{NW_TP}} + w_4 \cdot r^{\text{NW_LB}}. \quad (6.1)$$

All reward components are normalized against estimated baselines. Baselines are initialized based on the performance of the conventional heuristic and are subsequently updated during training. Each component is assigned a weight w_c with $c \in \{1, 2, 3, 4\}$ to control its relative importance within the overall reward structure.

$T_i^{\text{RL}}(t)$ denotes the total throughput of cell i under the learned policy, and $T^{\text{RL}}(t) = \sum_i T_i^{\text{RL}}(t)$ denotes the corresponding network-wide throughput. Furthermore, $\sigma^{\text{RL}}(t)$ denotes the standard deviation of per-cell throughputs. Using moving baseline estimators $B_i^{\text{Cell}}(t)$, $B_i^{\text{UE}}(t)$, $B^{\text{NW}}(t)$, and $B^\sigma(t)$, all

evaluated at the current decision step t , the reward components are defined as

$$r_i^{\text{Cell_TP}} = \frac{T_i^{\text{RL}}(t)}{B_i^{\text{Cell}}(t)} - 1, \quad (6.2)$$

$$r_i^{\text{UE_QoE}} = \frac{1}{|\mathcal{S}_i|} \sum_{u \in \mathcal{S}_i} \log \left(\frac{R_u^{\text{RL}} + \epsilon}{R'_u + \epsilon} \right) - B_i^{\text{UE}}(t), \quad (6.3)$$

$$r^{\text{NW_TP}} = \frac{T^{\text{RL}}(t)}{B^{\text{NW}}(t)} - 1, \quad (6.4)$$

$$r^{\text{NW_LB}} = \frac{B^\sigma(t) - \sigma^{\text{RL}}(t)}{B^\sigma(t) + \epsilon}. \quad (6.5)$$

The local baseline estimators are keyed by seed, decision step, and cell identifier, whereas the global baseline estimators are keyed only by seed and decision step. The reward structure incorporates two local components: the own-cell throughput $r_i^{\text{Cell_TP}}$ and the local neighborhood component $r_i^{\text{UE_QoE}}$. The term $r_i^{\text{Cell_TP}}$ measures the total improvement in UEs throughput for cell i relative to the baseline estimate. The term $r_i^{\text{UE_QoE}}$ quantifies the average relative throughput gain for all non-idle UEs u within cell i 's focal neighborhood. This focal neighborhood comprises cell i itself and up to $k = 3$ adjacent neighboring cells. The set $\mathcal{S}_i$ contains all non-idle UEs within the focal neighborhood under either the RL policy or the conventional heuristic baseline, where R_u^{RL} and R'_u denote the corresponding RL and conventional heuristic throughputs for each UE u . We employ a logarithmic ratio to reward relative rather than absolute improvement, which addresses the wide magnitude differences in UE throughputs. This formulation ensures that multiplicative gains matter more than absolute differences, treats gains and losses symmetrically, and grants stronger credit to the improvement of low-throughput UEs. Because $\mathcal{S}_i$ aggregates UEs from both RL and conventional heuristic outcomes, the agent cannot artificially increase the reward by merely moving poorly served UEs out of its neighborhood.

The reward formulation incorporates two shared global components: the total network throughput improvement $r^{\text{NW_TP}}$ and the per-cell throughput standard deviation reduction $r^{\text{NW_LB}}$. The term $r^{\text{NW_TP}}$ measures the improvement in total network throughput relative to the baseline estimate and rewards all agents jointly when throughput increases. The term $r^{\text{NW_LB}}$ captures load balancing by measuring the standard deviation of per-cell throughputs. Because lower standard deviation indi-

cates better load balancing, we define $r^{\text{NW_LB}}$ as a normalized reduction relative to the baseline, so that positive values denote improvement. These global terms consist of scalar indicators already measured by the network [11] that can be broadcast to all cells at the beginning of each round with negligible communication overhead. Overall, the reward function effectively balances local cell performance and coordination objectives while relying on minimal global information.

NEIGHBOR-SLOT LIMITATION. The neighborhood observation restricts each agent to at most the k strongest neighboring cells, with any missing slots zero-padded according to Table 6.2. This cap preserves a fixed per-agent observation dimension that remains independent of the physical neighborhood graph density. The constraint is important for scalability in denser deployments, where exposing all potential neighbors would render the input size topology-dependent.

POLICY CHECKPOINT SELECTION. Throughout training, we periodically store policy checkpoints containing the full set of per-agent models. Because these checkpoints yield different trade-offs among reward components, no single checkpoint necessarily dominates all objectives. For evaluation, we select checkpoints using a reproducible composite score computed over global throughput, global load balancing, and local UE QoE. The three metrics are assigned equal weights. Checkpoint selection is performed manually after training; automating this offline model-selection step is treated as an engineering extension rather than part of the learning formulation studied here. Section 6.3 presents and discusses the evolution of the reported metrics throughout distinct training processes. The resulting figures demonstrate that distinct training stages impose different trade-offs among the evaluated metrics. This temporal analysis motivates the checkpoint-selection procedure.

TRAINING STABILIZATION. To stabilize training and mitigate abrupt deviations from the initial operating point, we fix the initial conventional baseline value as a permanent anchor in the moving baseline estimator. Once populated, the baseline prediction becomes the mean of this fixed conventional heuristic anchor and the most recent RL sample(s) at the corresponding key. This anchor does not impose a hard limit on how far the learned policy can deviate from the conventional

heuristic. Rather, this mechanism stabilizes the learning dynamics by slowing abrupt changes in the reward baseline, thereby regulating the rate at which the policy adapts away from the initial known operating point. While this approach can in principle slow adaptation to rapidly improving policies, it stabilizes training especially for the load-balancing component, which we find to be particularly prone to drifting and instability without this preservation. This idea is conceptually related to common update-stabilization techniques such as adaptive optimization with AdamW [41] and gradient clipping [56]. However, those methods act on the parameter update itself, whereas the permanent anchor stabilizes the reward-normalization reference from which the policy-gradient signal is computed. It is therefore complementary rather than redundant, because it addresses instability caused by a rapidly moving reward baseline in a multi-seed, multi-objective training setting.

Figure 6.1 compares the training dynamics of configurations with and without the permanent anchor. While both configurations achieve substantial improvements across all three metrics relative to the conventional heuristic, their training dynamics differ notably. Without the permanent anchor, the global throughput gain converges earlier but saturates at a substantially lower level, while the load-balancing gain exhibits early improvement followed by performance drift and consistently lower values in later training stages. With the permanent anchor, later checkpoints maintain higher load-balancing gains, and global throughput continues to improve rather than saturating early. In contrast, the unanchored variant yields stronger local UE gains throughout training, yet incurs a penalty in global coordination. Overall, the permanent anchor steers training toward a more robust global trade-off, whereas the unanchored variant prioritizes the local UE objective. This behavior demonstrates that checkpoint selection must evaluate the complete metric trade-off on unseen seeds, rather than optimizing for a single training trajectory.

6.3 EVALUATION

This section evaluates the proposed formulation along three primary questions. First, we assess whether the learned policy improves upon the conventional heuristic approach. Second, we investigate whether the approach retains the gains of the prior centralized formulation under a matched slow-control setting.

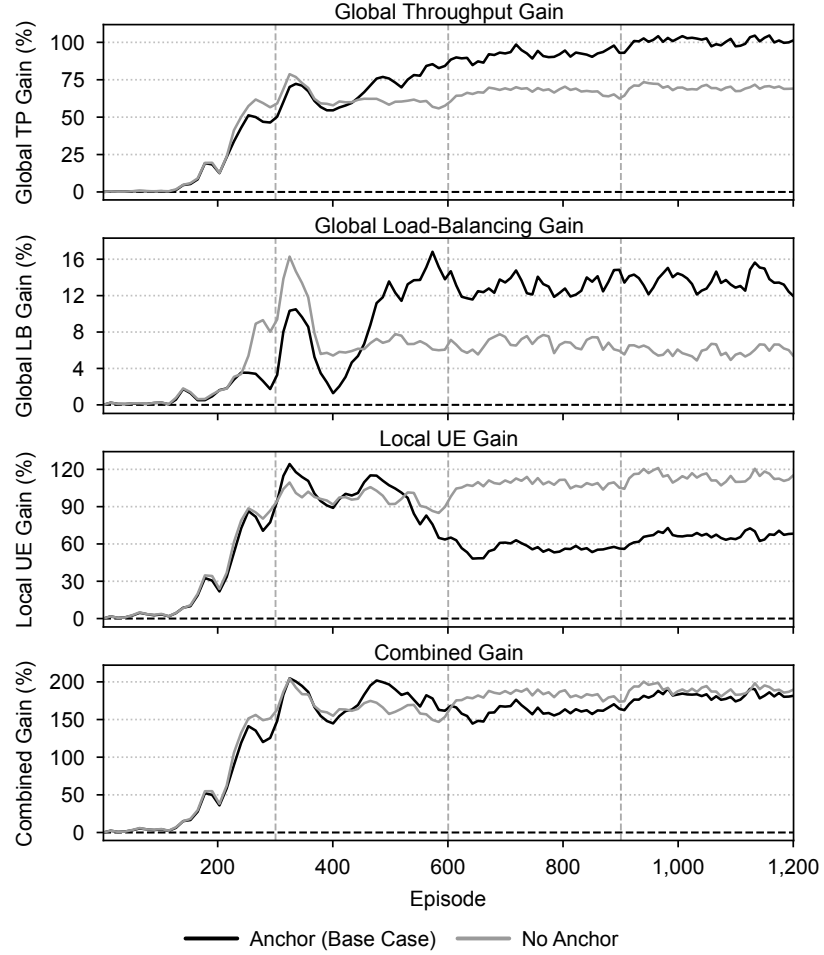


Figure 6.1: Effect of permanent baseline anchoring on multi-agent training dynamics. The curves compare the anchored base case against an unanchored variant across training episodes, reporting global throughput gain, global load-balancing gain, local UE gain, and the combined objective, demonstrating that the anchor improves global coordination at the expense of reduced local UE gain. Metrics are smoothed using a rolling mean over a 25-episode window. Reprinted from [35], © 2026 IEEE.

Finally, we examine whether these performance gains persist under deployment-oriented constraints and scenario shifts.

6.3.1 *Evaluation Setup*

The learning mechanisms follow the common framework established in Chapter 4; here, we specify the concrete parameter instantiations and configurations required for reproducible training.

We train the policy over 100 distinct episodes, where each episode employs a unique random seed to determine UE positions, mode transitions, mobility patterns, and other stochastic environmental aspects. Evaluation is conducted on 100 unseen seeds for each scenario to assess generalization performance across diverse random realizations. The control loop permits the agent to take an action at every simulation step, operates without UE mobility, and models mode transitions via rates $\lambda_{\text{IDLE}} = \lambda_{\text{ACTIVE}} = 0.2$. Training and evaluation typically let the agent update the parameters at every simulation step ($\text{UDR} < 1$), except in the 10-step control scenario, where updates occur every 10 simulation steps ($\text{UDR} > 1$). The neighborhood observation is constrained to at most $k = 3$ strongest neighboring cells per agent. Action space ranges are normalized according to Table 3.2. Exploration is managed using a Gaussian distribution with the standard deviation σ capped at 0.1 (see Figure 4.5). The base case utilizes the full observation space described in Section 6.1, a history length of $h = 10$, and four reward components with equal weights ($w_1 = w_2 = w_3 = w_4 = 0.25$). Each reward component maintains a separate baseline estimator, where global estimators are keyed only by decision step and local estimators are additionally keyed by cell-agent identifier. The baseline estimators employ a configurable window size w that determines the number of samples used in the moving average; we set $w = 2$ or $w = 3$ depending on the evaluation scenario (see Section 6.3). By default, all baseline estimators employ a permanent anchor derived from the conventional heuristic to stabilize the training dynamics and regulate the rate of policy adaptation. We configure the neural network architecture with 1024 neurons with 2 hidden layers per agent to balance sufficient policy capacity with computational tractability. We include an ablation study over alternative hidden-layer widths in Section 6.3.4 to assess sensitivity to neural network capacity. Even when deployed individually for each agent, the per-agent computational load

remains minimal, requiring fewer than 5 MFLOPs per inference pass and enabling sub-millisecond execution on commodity hardware. We apply a gradient clipping threshold of 10 to ensure training stability. The initial learning rate is set to 10^{-4} and is halved at each new curriculum-learning round, while the episode length starts at 30 steps and increases by 5 steps per round across 4 total rounds, resulting in a final episode length of 45 steps.

The default experiment employs a custom simulator, operator deployment data, and the common training and evaluation protocol. Subsequent scenarios are constructed by varying one factor at a time relative to the base configuration.

BASE-CASE PERFORMANCE. The conventional heuristic baseline employs the operator-derived parameters listed in Table 3.1, representing static expert-tuned cell (re)selection without adaptive learning or inter-cell coordination. Figure 6.2 illustrates the base-case gains of the proposed multi-agent policy relative to this reference. The policy achieves median gains of 133 % in global throughput, 11 % in global load balancing, and 94 % in local UE QoE. This baseline performance establishes the reference point for the subsequent robustness and generalization experiments, where one deployment condition is altered at a time.

COMPARISON WITH PRIOR CENTRALIZED FORMULATION. The evaluation compares the proposed approach against the prior centralized homogeneous formulation (see Chapter 5) under a matched experimental setup. The policy operates under an identical reduced-control setting, updating cell (re)selection parameters every 10 simulation steps. All reported gains are measured relative to the same conventional heuristic baseline. Under these matched conditions, the distributed cell-specific formulation achieves slightly higher median gains across all metrics: 144 % vs. 139 % in global throughput, 24 % vs. 20 % in global load balancing, and 202 % vs. 175 % in local UE QoE. This comparison indicates that shifting from centralized area-wide control to distributed per-cell control improves operational fit without sacrificing the performance benefits of learning-based cell (re)selection.

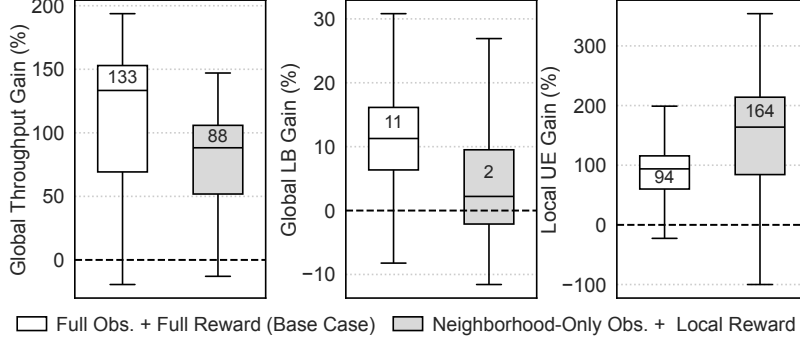


Figure 6.2: Effect of global observations and reward terms on multi-agent performance. The boxplots compare the full-observation, full-reward base case against a neighborhood-only variant with local reward terms, demonstrating that local information preserves strong throughput and local **UE** gains but noticeably reduces global load-balancing gains. Reprinted from [35], © 2026 IEEE.

6.3.2 Robustness Under Deployment-Oriented Constraints

This section evaluates policy effectiveness under practical deployment constraints. We assess three distinct operational dimensions: the availability of agent information, **UE** mobility, and the control update frequency.

NEIGHBORHOOD-ONLY INFORMATION. This configuration removes the shared global observation features while retaining only the two local reward terms ($w_1 = w_2 = 0.5$, $w_3 = w_4 = 0$). The modification reduces the per-agent input size from 130 to 80 for a history length of $h = 10$. We design this setup to test whether neighborhood-level information alone can support effective learning when global state features and global reward shaping are excluded. The resulting evaluation isolates the contribution of local coordination mechanisms under restricted information availability.

Figure 6.2 demonstrates that the neighborhood-only variant outperforms the conventional heuristic baseline. The policy achieves substantial global throughput gains alongside an even stronger local **UE** gain. However, the load-balancing gain drops sharply, indicating that this objective depends more directly on global coordination. The resulting wider gain distribution further suggests that removing global information reduces outcome consistency across random seeds. Thus, local and neighborhood information remains sufficient for strong local gains,

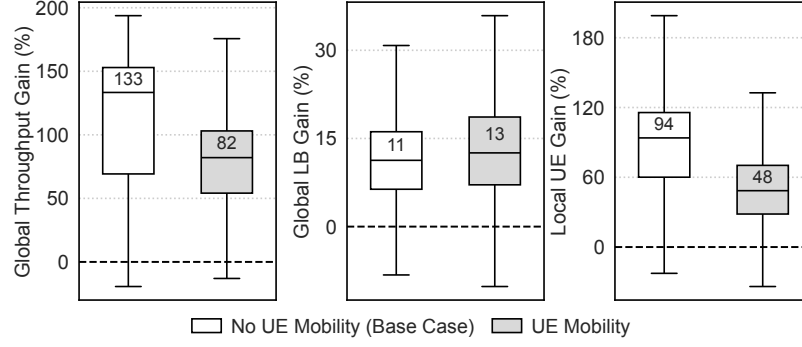


Figure 6.3: Effect of UE mobility on multi-agent performance. The boxplots compare the static base case against a mobility scenario with street-level UE movement during training and evaluation, demonstrating that mobility reduces global throughput and local UE gains while retaining global load-balancing gains. Reprinted from [35], © 2026 IEEE.

while shared global terms are necessary for balancing these improvements with network-wide coordination.

UE MOBILITY. We enable UE mobility during both training and evaluation to assess policy behavior under dynamic conditions. Users follow street-level trajectories at an average speed of 30 km/h.

Figure 6.3 demonstrates that the learned policy continues to outperform the conventional heuristic baseline under mobile UEs. Mobility reduces both global throughput and local UE gains relative to the static base case, indicating that changing UE positions makes sustained local improvements harder to maintain. In contrast, the retained load-balancing gain suggests that aggregate coordination remains robust even when UE-level optimization becomes more difficult.

The corresponding training dynamics for this configuration are shown in Figure 6.4. It compares the training dynamics of the base case against the mobility setting. Both configurations learn policies that remain clearly above the conventional heuristic baseline, demonstrating that UE mobility introduces additional difficulty without preventing stable learning. The primary effect of mobility is a reduction in global throughput and, more strongly, in local UE gain. In particular, the local UE gain rises early in training but subsequently settles at a lower level, indicating that sustained per-UE improvements are harder to maintain when users continuously move through the network.

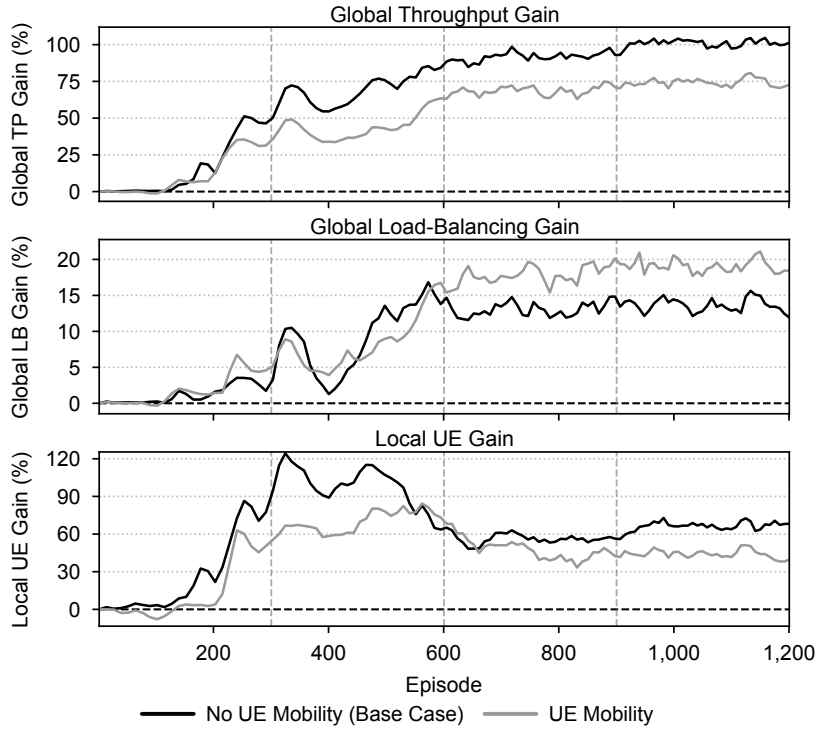


Figure 6.4: Multi-agent training dynamics with and without UE mobility. The curves compare the static base case against the mobility scenario across training episodes, reporting global throughput gain, global load-balancing gain, and local UE gain, demonstrating how mobility alters the trade-off between local and global objectives. Metrics are smoothed using a rolling mean over a 25-episode window. Reprinted from [35], © 2026 IEEE.

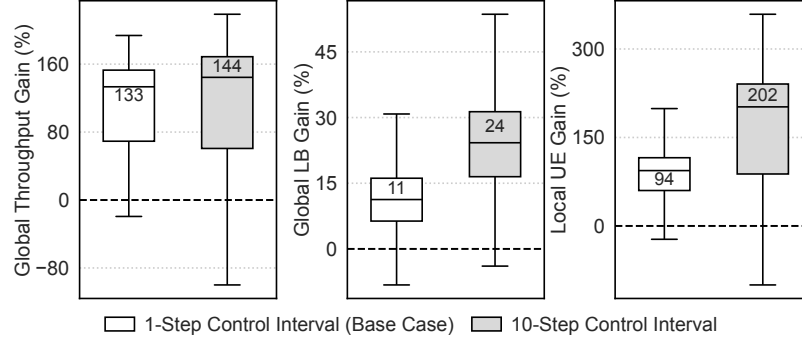


Figure 6.5: Effect of reduced control frequency on multi-agent performance. The boxplots compare the 1-step base case against a 10-step control interval, reporting global throughput gain, global load-balancing gain, and local UE gain, demonstrating that slower updates preserve or improve performance under the evaluated setting. Reprinted from [35], © 2026 IEEE.

However, the load-balancing gain remains robust under mobility and becomes even stronger in later training stages. A plausible explanation is that mobility naturally mixes UE populations across cells and creates more frequent reselection opportunities, which helps dissolve persistent cell overloads. Thus, mobility weakens fine-grained local UE optimization but preserves useful aggregate coordination behavior.

REDUCED CONTROL FREQUENCY. In this scenario, the controller is restricted to updating cell (re)selection parameters every 10 simulation steps. This design establishes a slower operational cadence that aligns with practical deployment constraints. The reduced update frequency prioritizes system stability and feasibility over aggressive real-time adaptation.

Figure 6.5 demonstrates that reducing the control frequency does not degrade performance, as the median gains remain at least as strong as those in the one-step base case. We observe analogous behavior in the single-agent setting in Section 5.2, confirming that the effect generalizes across agent counts. This configuration suggests that slower updates may function as a form of temporal regularization by mitigating overreaction to short-term fluctuations and permitting the network state to settle prior to the next action. Consequently, practical control delays may enhance both effectiveness and deployment feasibility by improving operational stability and coordination.



Figure 6.6: Multi-agent training dynamics under reduced control frequency. The curves compare the 1-step base case against a 10-step control interval across training episodes, reporting global throughput gain, global load-balancing gain, and local UE gain, demonstrating how slower updates influence convergence and objective trade-offs. Metrics are smoothed using a rolling mean over a 25-episode window. Reprinted from [35], © 2026 IEEE.

The corresponding training dynamics are shown in Figure 6.6. It compares the training dynamics of the base case against a configuration that updates parameters every 10 simulation steps. Both configurations learn policies that remain clearly above the conventional heuristic baseline. The 10-step control interval exhibits stronger and more stable load-balancing gains throughout much of the training process. Global throughput converges to comparable levels in both configurations, demonstrating that reduced control frequency does not impede the development of effective global coordination policies.

The primary distinction between the configurations manifests in the local UE and load-balancing terms. In the 1-step base case, the local UE gain peaks early and subsequently settles at a lower level as training shifts toward a more balanced global operating point. Conversely, the 10-step update interval maintains a higher local UE gain in later training stages while load balancing improves substantially. This aligns with our earlier observations regarding control frequency and network stability.

6.3.3 Generalization Across Topology, Load, and Seed Diversity

This section evaluates the effectiveness of the learned policy beyond the baseline training setting. The analysis considers deliberate changes to the deployment topology, network load conditions, and training-set diversity. These experiments assess the formulation sensitivity to scenario shifts and broader training exposure.

GENERALIZATION TO OTHER LOCATIONS. This experiment assesses the formulation dependence on the specific baseline topology. Because the distributed architecture assigns a cell-specific policy to each agent, direct policy transfer across deployments is infeasible; each policy is locally fine-tuned to its corresponding cell. We therefore retrain and evaluate the formulation from scratch in a geographically distinct area (see Figure 4.2) containing the same number of towers and cells but featuring different neighborhood relations. Figure 6.7 shows the neighborhood graph, where each node corresponds to an agent controlling exactly one cell. It illustrates the resulting topology shift between the baseline and alternative deployment areas. The baseline topology comprises two dense neighborhoods, one sparse neighborhood (5 nodes), and three isolated nodes, whereas the alternative topology features one dense

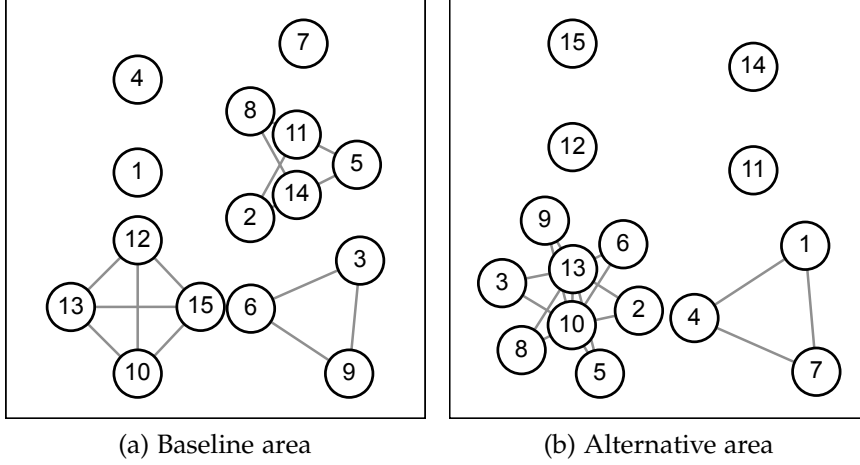


Figure 6.7: Neighborhood graphs for the multi-agent formulation in the baseline and alternative areas. Nodes represent cells, and edges denote precomputed neighbor relations derived from coverage overlap and signal-strength checks. The distinct graph structures illustrate the topology shift used for the geographic generalization experiment. Reprinted from [35], © 2026 IEEE.

neighborhood, one larger sparse neighborhood (8 nodes), and four isolated nodes. This topological shift implies that while dense neighborhoods can leverage high local coordination density, the larger sparse cluster in the alternative topology relies more strongly on the compact global aggregate signaling to maintain coordination.

Figure 6.8 demonstrates that the proposed approach remains clearly above the conventional heuristic in the alternative area. The global throughput gain is reduced compared to the baseline area, indicating that the exact topology and propagation conditions affect the achievable optimum. However, load balancing and local UE performance remain strong. These results support the conclusion that the neighborhood-aware formulation is not specific to one deployment graph.

GENERALIZATION ACROSS NETWORK LOAD LEVELS. We train the policy with 500 UEs to evaluate load generalization. We subsequently assess the same policy under lower (100) and higher (1000) UEs load conditions. Figure 6.9 shows that the learned policy consistently outperforms the conventional heuristic across all evaluated load levels. The performance gains are smaller at lower load due to reduced congestion and limited

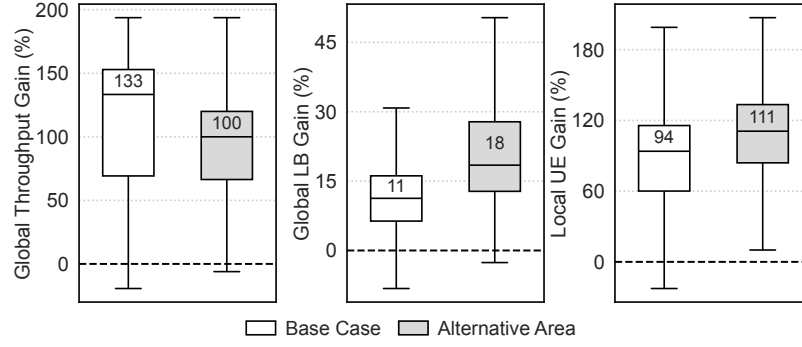


Figure 6.8: Multi-agent performance gains relative to the conventional heuristic in the baseline and alternative areas. The boxplots compare global throughput, global load balancing, and local UE gains, demonstrating that the learning formulation remains effective under a geographically distinct deployment topology. Reprinted from [35], © 2026 IEEE.

opportunities for beneficial cell (re)selection and load balancing. Conversely, the gains become substantially stronger at the training load and under higher load, indicating that the policy maintains its effectiveness beyond the specific load regime used during training. These results demonstrate reliable generalization across varying traffic densities.

INCREASED TRAINING DIVERSITY. We evaluate the impact of increased training diversity by expanding the training set from 100 to 500 seeds, which yields 500 unique episodes. The evaluation is subsequently conducted on 100 unseen seeds to isolate the effect of training distribution on policy generalization.

Figure 6.10 shows that broader training diversity preserves clear gains over the conventional heuristic while altering the metric trade-off. The local UE gain increases, whereas global throughput and load-balancing gains decrease moderately. This shift indicates that exposing the policy to a wider range of scenarios directs learning toward the denser local UE signal and reduces specialization to high-gain global coordination patterns.

Figure 6.11 presents the corresponding training dynamics to illustrate this behavioral transition. The base case training dynamics are omitted from this figure because the expanded scenario diversity alters the number of episodes, making direct visual comparison of training trajectories difficult. The broader training distribution generates strong early gains, particularly for global throughput and local UE performance. However, the resulting

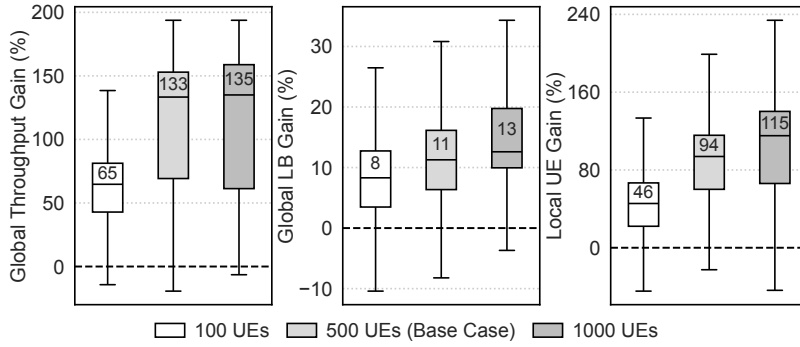


Figure 6.9: Multi-agent performance gains relative to the conventional heuristic across different UE load levels. The policy is trained on 500 UEs and evaluated with 100, 500, and 1000 UEs, demonstrating improved performance in global throughput, global load balancing, and local UE metrics across the evaluated load levels. Reprinted from [35], © 2026 IEEE.

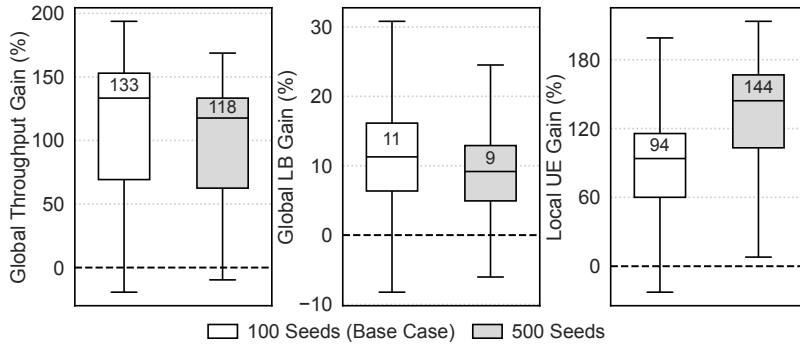


Figure 6.10: Effect of increased training diversity on multi-agent performance relative to the conventional heuristic. The box-plots compare the 100-seed base case against a policy trained with 500 seeds, demonstrating that broader training diversity increases local UE gains while moderately reducing global throughput and load-balancing gains. Reprinted from [35], © 2026 IEEE.

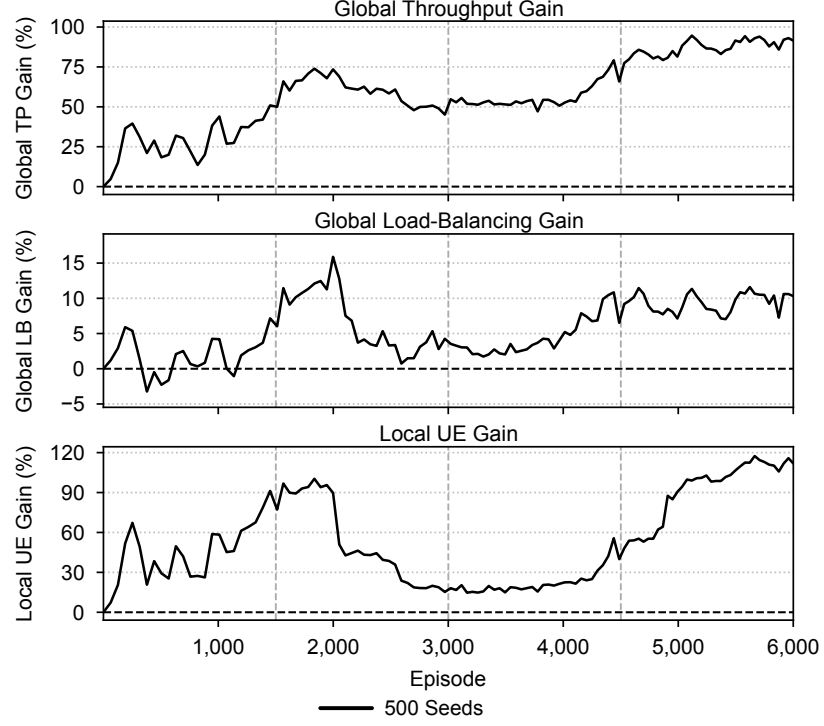


Figure 6.11: Multi-agent training dynamics under increased training diversity with 500 seeds. The curves show global throughput gain, global load-balancing gain, and local UE gain across training episodes, demonstrating the extended and more variable optimization process resulting from the larger training set. Metrics are smoothed using a rolling mean over a 25-episode window. Reprinted from [35], © 2026 IEEE.

curves are less monotonic than those of the base case. Following the initial improvement, both local UE and load-balancing gains decline during the middle of training before recovering in the later curriculum stages.

This behavior indicates that broader training diversity makes the optimization problem less specialized and shifts the metric trade-off throughout training. In the final phase, global throughput and local UE gains recover to strong levels, while load balancing remains positive but more modest. This dynamic supports the checkpoint-selection procedure used in the main evaluation: under broader scenario diversity, the best checkpoint is not necessarily the earliest high-gain point, but the one that yields the optimal trade-off across the reported metrics.

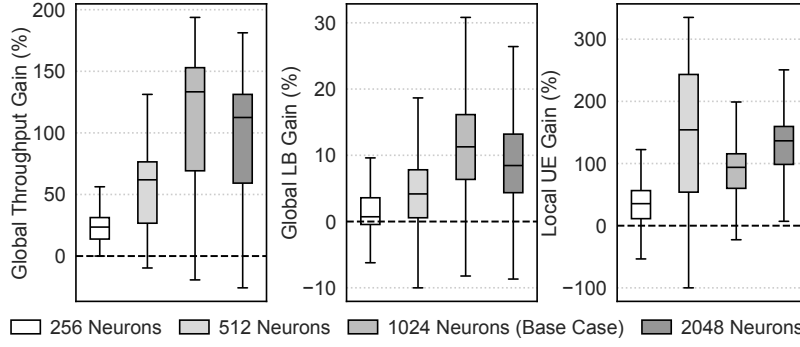


Figure 6.12: Effect of per-agent policy network size on multi-agent performance. The boxplots compare hidden-layer widths of 256, 512, 1024, and 2048 neurons, reporting global throughput gain, global load-balancing gain, and local UE gain, demonstrating that the 1024-neuron base architecture provides the strongest overall trade-off. Reprinted from [35], © 2026 IEEE.

6.3.4 Design Choices and Architecture Sensitivity

While the preceding scenarios evaluate performance under explicit constraint shifts, we further validate the reported gains by examining the impact of implementation choices. This section assesses sensitivity to model capacity, reward-shaping parameters, and observation design to confirm that the base configuration represents the optimal trade-off across stability, complexity, and deployment constraints.

NEURAL NETWORK SIZE. This subsection presents an ablation study on the neural network size to evaluate the impact of model capacity on the learned policy. We compare hidden-layer widths of 256, 512, 1024 (base), and 2048 neurons across the reported performance metrics. The corresponding results are detailed in Figure 6.12 and Table 6.3. The analysis confirms that the 1024-neuron configuration achieves the strongest overall balance across metrics, validating its use as the base architecture.

The results reveal a clear capacity effect as the hidden-layer width increases from 256 to 1024 neurons. Configurations with 256 neurons exhibit clear underfitting across all three evaluated metrics. Increasing the width to 512 neurons achieves strong local-UE optimization but yields weaker global coordination and a large spread in local UE gain, indicating strong but less consistent user-level improvements across seeds. The 1024-neuron configuration achieves the strongest overall balance across all

Table 6.3: Median gains (%) over the conventional heuristic across hidden-layer widths

Width	Global TP	Global LB	Local UE
256	23.5	0.7	35.4
512	61.9	4.2	154.2
1024 (base)	133.3	11.3	93.9
2048	112.5	8.5	136.5

metrics. Expanding the width to 2048 neurons provides no further performance improvement, demonstrating that a larger model is not automatically better. These findings support the choice of a lightweight architecture, showing that very large networks are not required to obtain strong gains. The base configuration requires less than 5 MFLOPs per inference pass, enabling sub-millisecond execution on commodity hardware and ensuring compatibility with fast practical deployment in real-world cellular networks.

BASELINE ESTIMATOR WINDOW SIZE. The moving reward baseline employs a configurable window size w that determines the number of recent samples included in the baseline estimate for a given seed, step, and reward key. Increasing this window smooths short-term reward fluctuations, but reduces the baseline responsiveness to recent policy changes.

Figure 6.13 demonstrates that increasing the baseline window from $w = 2$ to $w = 3$ alters the metric trade-off rather than uniformly improving all objectives. A larger window improves local UE gain, but consequently reduces both global throughput and load-balancing gain. This behavior suggests that the smoother, slower-adapting baseline aligns better with the noisier local UE objective, while simultaneously weakening responsiveness to changes in the global reward structure. The base-case choice of $w = 2$ therefore maintains a superior balance across the three reported metrics.

NEIGHBORHOOD HISTOGRAM AUGMENTATION. We evaluate an augmented observation design that adds a compact radio-quality profile for each cell represented in the local neighborhood observation. For the agent’s cell and each configured neighbor slot, the observation includes a normalized histogram of RSRP values from currently connected UEs, binned into fixed

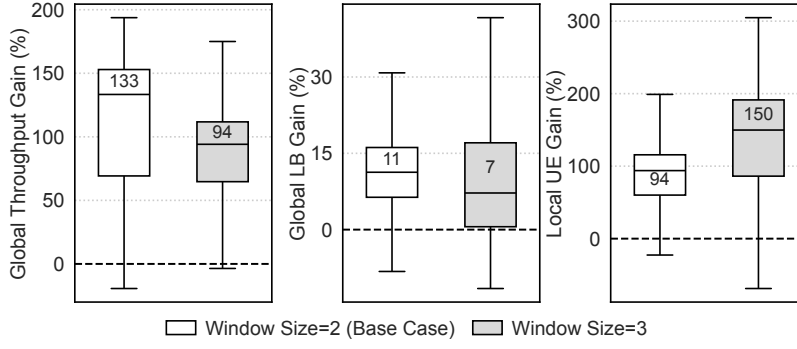


Figure 6.13: Effect of the baseline estimator window size on multi-agent performance. The boxplots compare the base configuration with window size $w = 2$ against a larger window size $w = 3$, reporting global throughput gain, global load-balancing gain, and local UE gain, demonstrating that a larger window shifts the trade-off toward stronger local UE gains. Reprinted from [35], © 2026 IEEE.

signal-strength intervals. The augmentation retains the fixed-size, relative observation design because the bin count is fixed and the bin values represent normalized fractions rather than absolute UE counts. It follows the same design principle as the neighborhood observation: representing variable-size radio and topology information through fixed-size summaries. This design provides a local summary of whether the observed cells mainly serve weak edge users or stronger center users, without exposing individual UE-level measurements. Missing neighbors or cells lacking active UEs are encoded as zero-valued histograms.

The histogram augmentation does not uniformly improve performance, as shown in Figure 6.14. The slight improvement in load balancing indicates that the additional radio-quality context enables the identification of redistribution opportunities across neighboring cells. However, both global throughput and local UE gains are reduced relative to the compact base observation. This suggests that incorporating spatial radio context shifts the learned trade-off and may increase input complexity without providing sufficient benefit for the throughput and UE objectives. Overall, the compact base observation yields the most balanced design.

6.3.5 Analysis of UE QoE Improvements

Aggregate summary metrics capture average performance across random seeds, but do not reveal whether improvements are

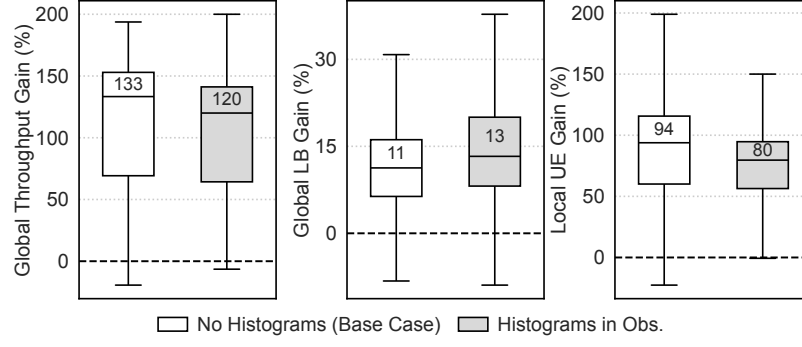


Figure 6.14: Effect of radio-quality histogram features on multi-agent performance. The boxplots compare the base observation against an augmented observation using normalized *RSRP* histograms for the agent cell and neighboring cells, demonstrating a modest load-balancing gain at the expense of reduced global throughput and local *UE* gains. Reprinted from [35], © 2026 IEEE.

concentrated in outliers or broadly distributed. We therefore analyze the full throughput distribution to validate the reward design and confirm that gains reflect consistent improvements across the user population.

We analyze the *UE QoE* metric by plotting the full distribution of *UEs* throughputs experienced under the learned policy and the conventional heuristic, as shown in Figure 6.15. Unlike step-wise boxplots, these cumulative distribution functions reveal how the policies differ across all experienced throughput values rather than solely capturing average improvements. The global *ECDF* aggregates all *UEs* throughputs over the evaluated episodes, while the neighborhood *ECDFs* aggregate throughputs for users connected to specific cells within each neighborhood or cluster. The global *ECDF* shows that the *RL* curve is shifted almost entirely to the right of the conventional heuristic, indicating broadly improved throughput across most of the distribution, particularly in the low-to-medium regime. The neighborhood-level *ECDFs* demonstrate that these *QoE* gains are also visible within individual local interaction structures, with subplots A and C showing that the *RL* policy consistently dominates the heuristic across almost the entire distribution. Subplot B exhibits a small overlap only in the lower-throughput regime, where the conventional heuristic is marginally better for a small fraction of samples, but the *RL* policy again shifts the distribution toward higher throughput outside this overlap. Three isolated cells without any neighbors are omitted from the neighborhood

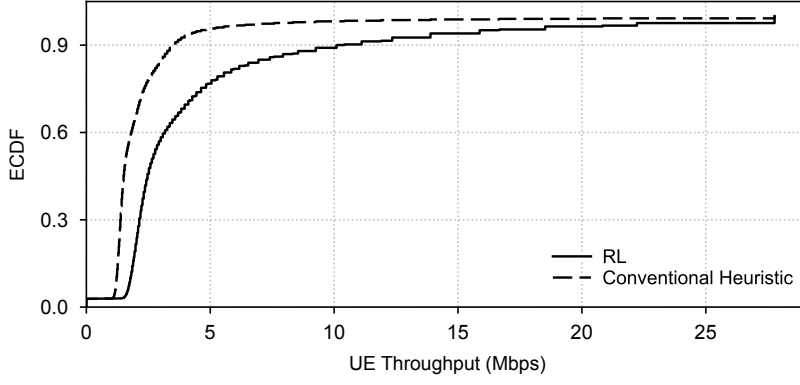


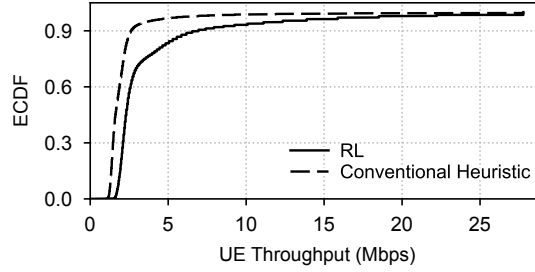
Figure 6.15: Global ECDF of UE throughput across all evaluated episodes. The learned RL policy shifts the distribution toward higher throughput relative to the conventional heuristic, indicating broadly improved UE QoE rather than gains restricted to a small subset of UEs. Reprinted from [35], © 2026 IEEE.

subplots. For these isolated cells, the RL policy shows an even stronger shift over the conventional heuristic, which is notable because the corresponding agents lack neighbor information and rely solely on their local state and the shared global summary. This observation suggests that the learned policy can still improve strongly when the decision problem is predominantly local. Overall, the ECDF analysis confirms that the learned policy broadly improves UE QoE across the experienced throughput distribution rather than relying on a few high-throughput outliers. This indicates robustness across different local neighborhood structures, which is essential for generalization to deployment scenarios with varying cell densities and neighbor configurations.

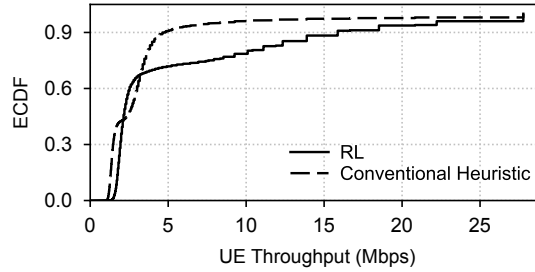
6.4 CONCLUSION

This chapter presented a distributed, cell-specific multi-agent formulation for adaptive idle-mode cell (re)selection. In contrast to the prior centralized homogeneous approach, each cell agent independently controls its own reselection parameters from local and neighborhood observations, while compact aggregate feedback supports network-wide coordination. This design preserves fixed per-agent action and observation spaces, aligning learning-based reselection control with scalable cell-level operation. Realistic simulations based on operator deployment data demonstrate that the proposed formulation consistently

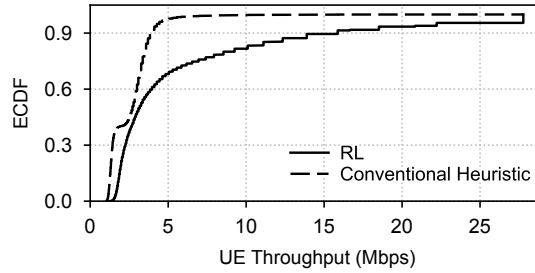
outperforms the conventional heuristic baseline and retains the primary performance gains observed with prior centralized RL-based control. Under a matched slow-control setting, the distributed formulation achieves comparable or slightly stronger median gains relative to the centralized approach. The system remains effective across diverse operational conditions, including neighborhood-only information, UE mobility, reduced control frequency, changed deployment topology, varying load levels, and increased training diversity. These results indicate that adaptive cell (re)selection can transition from centralized proof-of-concept designs toward practical distributed control architectures while preserving the core benefits of learning-based parameter optimization.



(a) Neighborhood nodes 3, 6 and 9



(b) Neighborhood nodes 10, 12, 13 and 15



(c) Neighborhood nodes 2, 5, 8, 11 and 14

Figure 6.16: Neighborhood-level ECDFs of UE throughput across evaluated episodes. Each subfigure aggregates UEs belonging to a single local neighborhood in the baseline area and compares the learned RL policy against the conventional heuristic, demonstrating that the throughput improvements observed globally also manifest within individual neighborhood structures. Adapted from [35], © 2026 IEEE.

CONCLUSION

This chapter synthesizes the energy-aware band switching framework and the adaptive cell (re)selection, evaluates their operational limitations and deployment constraints, and outlines a roadmap for standardization and future joint optimization research.

7.1 CONNECTED-MODE SYNTHESIS

This synthesis evaluates the transition from conventional signal-strength-based handovers to energy-aware band switching, validating substantial power savings through offline optimization.

ENERGY-FIRST BAND SELECTION: REDEFINING CONNECTED-MODE OBJECTIVES. The proposed energy-aware band switching framework establishes a rigorous foundation for optimizing **UE** power consumption across heterogeneous 5G deployments. By deriving an analytical power model and calibrating it against commercial measurements, the system-level abstraction successfully bridges slot-level activities to macro-scale predictions, capturing non-linear relationships between allocated bandwidth, traffic load, and transmission power. The offline **MIQCP** optimization framework validates this model, demonstrating that energy-aware handover decisions can reduce total **UE** energy consumption by up to 25.5 % relative to conventional **RSRP**-centric baselines while preserving fairness and meeting application data rate constraints. Crucially, the results reveal that narrow-band, high-power configurations can outperform wide-band, low-power setups beyond specific packet rates, exposing the fundamental inefficiency of signal-strength-only handover criteria. These findings confirm that power consumption must be treated as a first-class optimization metric rather than a secondary byproduct of mobility management.

LIMITATIONS OF STATIC OFFLINE PLANNING. Despite these validated energy gains, the current formulation faces inherent operational constraints that limit direct deployment. The offline optimization paradigm requires complete knowledge

of future mobility patterns, traffic forecasts, and pre-computed rate/energy tables, thereby limiting its applicability in real-time network environments. Furthermore, event-triggered switching inherently incurs exploration overhead; continuous band probing introduces energy penalties that must be explicitly modeled to avoid negating optimization gains. Additionally, the requirement of real-time mobility data and accurate predictive models creates a significant barrier to online implementation. These limitations underscore the computational and data-driven challenges that static optimization frameworks cannot resolve, necessitating a shift toward adaptive, learning-based control mechanisms capable of operating under partial observability and dynamic network conditions.

7.2 IDLE-MODE SYNTHESIS

This synthesis traces the architectural evolution of idle-mode cell (re)selection from static operator heuristics to distributed multi-agent [RL](#).

LEARNING DYNAMICS AND POLICY BEHAVIOR. The proposed control framework evolves through three distinct operational paradigms: static operator-tuned heuristics, centralized single-agent [RL](#), and distributed multi-agent coordination. Initially, network operators rely on fixed, experience-driven parameter sets that operate independently of real-time conditions. This static baseline is superseded by CellPilot, a single-agent [RL](#) formulation that optimizes area-wide parameters through global observations. However, homogeneous parameterization imposes a global compromise that inadequately addresses heterogeneous local conditions. The final evolution assigns one decision-making agent per cell, enabling distributed, cell-specific parameter control while preserving network-wide coordination through neighborhood-aware observations and shared reward terms. This architectural progression resolves scalability bottlenecks and aligns learning-based control with practical deployment constraints.

Convergence dynamics differ markedly between the single- and multi-agent formulations. The single-agent policy converges toward global network aggregates, but naturally yields disproportionately large throughput gains in certain scenarios. To account for this and maintain balanced optimization across all metrics, reward weights are explicitly tuned, ensuring that load

balancing remains optimized alongside throughput. In contrast, the multi-agent formulation exhibits distinct training trajectories where per-cell agents balance local QoE improvements with neighborhood-level coordination. Credit assignment in the distributed setting is achieved without omniscient control through compact, fixed-size neighborhood and global aggregates. As demonstrated in the training dynamics, the multi-agent formulation explicitly incorporates a localized UE QoE term into its reward structure. This design prioritizes neighborhood-level user experience while preserving robust global throughput and load-balancing performance. The permanent baseline anchor further stabilizes these dynamics by preventing abrupt deviations from the conventional heuristic operating point, ensuring that later checkpoints maintain stronger load-balancing gains without sacrificing global coordination.

DEPLOYMENT REALITIES AND STANDARDS INTEGRATION. The proposed control framework maps continuous RL outputs directly to 3GPP-specified idle-mode procedures, ensuring seamless integration with existing network architectures. The policy outputs continuous reselection parameters (e.g., hysteresis, thresholds, offsets) that align with SIB broadcasting mechanisms and operator-configurable thresholds defined in 3GPP specifications. By adapting these parameters in real time, the framework replaces static, experience-driven workflows with data-driven optimization while remaining fully compatible with current 3GPP standards for idle-mode cell (re)selection.

Hardware feasibility is validated through computational profiling of the proposed architecture. The compact four-layer feed-forward network requires fewer than 5 MFLOPs per inference pass, enabling sub-millisecond execution on commodity RAN hardware. This lightweight design, combined with a flat normalized input vector and efficient policy evaluation, ensures that real-time parameter adaptation can be deployed without imposing prohibitive computational burdens on existing base station infrastructure or RICs.

Signaling overhead is quantified by restricting inter-agent communication to fixed-size neighborhood exchanges comprising ≤ 13 features per decision round. This bounded information exchange aligns with existing X2 and E2 interface constraints. By transmitting only local summaries and compact global aggregates, the framework maintains low communication overhead

while preserving sufficient context for coordinated decision-making, making it viable for dense multi-cell deployments.

LIMITATIONS, OPEN CHALLENGES AND FUTURE WORK. The current formulation intentionally excludes cell priorities from the action space, fixing them to conventional heuristic baselines to avoid discrete, high-impact reselection shifts. While this simplifies the continuous action space and stabilizes training, dynamic priority assignment remains unexplored. Future work must investigate adaptive priority schemes that respond to real-time congestion and application demands without destabilizing the reselection logic or violating 3GPP scheduling constraints.

While simulations leverage operator-derived topologies and commercial 5G measurements, the framework lacks real-world pilot testing, hardware-in-the-loop validation, and long-term drift monitoring. Translating learned policies to production networks requires rigorous field trials to assess robustness under unpredictable environmental conditions. Bridging this gap demands standardized testbeds, phased deployment strategies, and continuous performance monitoring.

A clear standardization pathway is required to transition from research prototypes to production-ready systems. This involves mapping [RL](#) outputs to 3GPP Self-Organizing Network ([SON](#)) frameworks, defining RAN Intelligent Controller integration patterns, and establishing benchmark suites for adaptive mobility management. Standardizing these interfaces will enable interoperable, vendor-agnostic deployment of [RL](#)-driven connectivity management across heterogeneous networks.

7.3 CONCLUDING REMARKS

While this thesis addresses connectivity management in **CONNECTED** and **IDLE** modes as complementary optimization domains, each independently delivers substantial practical value for real-world deployment. The energy-aware band switching framework and the distributed multi-agent cell (re)selection policy provide validated, standards-compatible mechanisms that operators can integrate immediately. Future work may explore joint simulation environments to model cross-[RRC](#) state transitions, developing coordination mechanisms that align handover triggers with reselection parameters. Such extensions

would enable seamless lifecycle optimization, building directly upon the independently validated components presented here.

BIBLIOGRAPHY

- [1] Iulian Rosu (VA3IUL). *Understanding Noise Figure*. 2023. URL: <https://www.qsl.net/va3iul/Noise/Understanding%20Noise%20Figure.pdf> (visited on 06/30/2026).
- [2] 3GPP. *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification*. ETSI Technical Specification ETSI TS 136 331 V8.2.0; 3GPP TS 36.331 version 8.2.0 Release 8. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI), Nov. 2008. URL: https://www.etsi.org/deliver/etsi_ts/136300_136399/136331/08.02.00_60/ts_136331v080200p.pdf (visited on 06/30/2026).
- [3] 3GPP. *Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects*. Technical Report TR 36.814 v9.0.0. 3GPP, Mar. 2010. URL: http://www.3gpp.org/ftp/Specs/archive/36_series/36.814/36814-900.zip (visited on 06/30/2026).
- [4] 3GPP. *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 application protocol (X2AP)*. Technical Specification TS 36.423 V8.9.0. Release 8. European Telecommunications Standards Institute (ETSI), Mar. 2010. URL: https://www.etsi.org/deliver/etsi_ts/136400_136499/136423/08.09.00_60/ts_136423v080900p.pdf (visited on 06/30/2026).
- [5] 3GPP. *Telecommunication management; Self-Organizing Networks (SON) Policy Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)*. Technical Specification 3GPP TS 28.628 V11.4.0. Release 11. European Telecommunications Standards Institute (ETSI), June 2014. URL: https://www.etsi.org/deliver/etsi_ts/128600_128699/128628/11.04.00_60/ts_128628v110400p.pdf (visited on 06/30/2026).
- [6] 3GPP. *Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; Vocabulary for 3GPP Specifications*. Technical Report TR 21.905 V14.1.1. European Telecommunications Standards Institute (ETSI), July 2017. URL: <https://www.etsi.org/d>

- [eliver/etsi_tr/121900_121999/121905/14.01.01_60/tr_121905v140101p.pdf](https://www.etsi.org/deliver/etsi_tr/121900_121999/121905/14.01.01_60/tr_121905v140101p.pdf) (visited on 06/30/2026).
- [7] 3GPP. *5G; NR; User Equipment (UE) procedures in idle mode and in RRC Inactive state*. Technical Specification TS 38.304 version 15.1.0 Release 15. European Telecommunications Standards Institute (ETSI), Oct. 2018. URL: https://www.etsi.org/deliver/etsi_ts/138300_138399/138304/15.01.00_60/ts_138304v150100p.pdf (visited on 06/30/2026).
 - [8] 3GPP. *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode*. Technical Specification TS 36.304 version 15.3.0 Release 15. European Telecommunications Standards Institute (ETSI), May 2019. URL: https://www.etsi.org/deliver/etsi_ts/136300_136399/136304/15.03.00_60/ts_136304v150300p.pdf (visited on 06/30/2026).
 - [9] 3GPP. *Study on User Equipment (UE) power saving in NR*. Technical Report TR 38.840 v16.0.0. 3GPP, June 2019. URL: <https://atis.org.s3.amazonaws.com/archive/3gpp-documents/Rel16/ATIS.3GPP.38.840.V1600.pdf> (visited on 06/30/2026).
 - [10] 3GPP. *5G; NR; Physical layer procedures for data*. Technical Specification ETSI TS 138 214 V16.2.0 / 3GPP TS 38.214 version 16.2.0 Release 16. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI), July 2020. URL: https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/16.02.00_60/ts_138214v160200p.pdf (visited on 06/30/2026).
 - [11] 3GPP. *5G; Management and orchestration; 5G performance measurements*. ETSI Technical Specification ETSI TS 128 552 V16.10.0; 3GPP TS 28.552 version 16.10.0 Release 16. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI), Aug. 2021. URL: https://www.etsi.org/deliver/etsi_ts/128500_128599/128552/16.10.00_60/ts_128552v161000p.pdf (visited on 06/30/2026).
 - [12] 3GPP. *5G; NR; Physical Channels and Modulation*. Technical Specification ETSI TS 138 211. 3GPP TS 38.211 version 16.10.0, Release 16. European Telecommunications Standards Institute (ETSI), July 2022. URL: https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/16.10.00_60/ts_138211v161000p.pdf (visited on 06/30/2026).

- [13] 3GPP. 5G; NR; *Radio Resource Control (RRC); Protocol specification*. ETSI Technical Specification ETSI TS 138 331 V17.1.0; 3GPP TS 38.331 version 17.1.0 Release 17. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI), Aug. 2022. URL: https://www.etsi.org/deliver/etsi_ts/138300_138399/138331/17.01.00_60/ts_138331v170100p.pdf (visited on 06/30/2026).
- [14] 3GPP. 5G; NR; *Radio Resource Control (RRC); Protocol specification*. Technical Specification TS 38.331 v15.16.0. 3GPP, Jan. 2022. URL: https://www.etsi.org/deliver/etsi_ts/138300_138399/138331/15.16.00_60/ts_138331v151600p.pdf (visited on 06/30/2026).
- [15] 3GPP. 5G; NR; *User Equipment (UE) procedures in idle mode and in RRC Inactive state*. Technical Specification ETSI TS 138 304 V17.0.0; 3GPP TS 38.304 version 17.0.0 Release 17. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI), May 2022. URL: https://www.etsi.org/deliver/etsi_ts/138300_138399/138304/17.00.00_60/ts_138304v170000p.pdf (visited on 06/30/2026).
- [16] 3GPP. 5G; NR; *Medium Access Control (MAC) protocol specification*. Technical Specification TS 38.321 v16.16.0. 3GPP, July 2024. URL: https://www.etsi.org/deliver/etsi_ts/138300_138399/138321/16.16.00_60/ts_138321v161600p.pdf (visited on 06/30/2026).
- [17] 3GPP. 5G; NR; *NR and NG-RAN Overall Description; Stage-2*. Technical Specification ETSI TS 138 300. 3GPP TS 38.300 version 18.1.0, Release 18. European Telecommunications Standards Institute (ETSI), May 2024. URL: https://www.etsi.org/deliver/etsi_ts/138300_138399/138300/18.01.00_60/ts_138300v180100p.pdf (visited on 06/30/2026).
- [18] Kamil Bechta, Christophe Grangeat, and Jinfeng Du. “Impact of Effective Antenna Pattern on Radio Frequency Exposure Evaluation for 5G Base Station with Directional Antennas.” In: *2020 XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science*. 2020. DOI: [10.23919/URSIGASS49373.2020.9232371](https://doi.org/10.23919/URSIGASS49373.2020.9232371).
- [19] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. “Curriculum learning.” In: *ACM ICML*. Vol. 382. 2009.

- [20] Bundesnetzagentur. *Bundesnetzagentur extends mobile communications map to include high-performance 5G standalone*. 2024. URL: https://www.bundesnetzagentur.de/SharedDocs/Pressemitteilungen/EN/2024/20240613_MoFu.html (visited on 06/13/2024).
- [21] CEPT. *CEPT Report 67: Report A from CEPT to the European Commission in response to the Mandate to develop harmonised technical conditions for spectrum use in support of the introduction of next-generation (5G) terrestrial wireless systems in the Union*. Review of the harmonised technical conditions applicable to the 3.4–3.8 GHz (‘3.6 GHz’) frequency band. Tech. rep. CEPT Report 67. European Conference of Postal and Telecommunications Administrations (CEPT), Electronic Communications Committee (ECC), July 2018. URL: <https://docdb.cept.org/download/118> (visited on 06/30/2026).
- [22] Kwangrok Chang, Ragil Putro Wicaksono, Seiji Kunishige, and Noriteru Takagaki. “LTE Idle Mode Optimization Improving End User Experiences.” In: *Proceedings of the 16th International Conference on Advanced Communication Technology (ICACT)*. Feb. 2014, pp. 14–18. ISBN: 978-89-968650-3-2. DOI: [10.1109/ICACT.2014.6778913](https://doi.org/10.1109/ICACT.2014.6778913).
- [23] ETSI/O-RAN Alliance. *E2 interface: General Aspects and Principles (O-RAN.WG3.E2GAP-R003-v04.01)*. Technical Specification ETSI TS 104 038 V4.1.0. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI), Oct. 2024. URL: https://www.etsi.org/deliver/etsi_ts/104000_104099/104038/04.01.00_60/ts_104038v040100p.pdf (visited on 06/30/2026).
- [24] Mohamed Fadl. *LTE Parameter Engineering Guidelines*. 2025. URL: <https://www.scribd.com/document/503090909/LTE-Parameter-Engineering-guidelines> (visited on 06/30/2026).
- [25] Rostand A. K. Fezeu, Claudio Fiandrino, Eman Ramadan, Jason Carpenter, Lilian Coelho de Freitas, Faaïq Bilal, Wei Ye, Joerg Widmer, Feng Qian, and Zhi-Li Zhang. “Unveiling the 5G Mid-Band Landscape: From Network Deployment to Performance and Application QoE.” In: *Proceedings of the ACM SIGCOMM 2024 Conference* (Sydney, NSW, Australia, Aug. 4–8, 2024). SIGCOMM ’24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 358–372. DOI: [10.1145/3651890.3672269](https://doi.org/10.1145/3651890.3672269).

- [26] Agrim Gupta, Adel Heidari, Avyakta Kalipattapu, Ish Kumar Jain, and Dinesh Bharadia. “3 W’s of smartphone power consumption: Who, Where and How much is draining my battery?” In: *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington, DC, USA, Nov. 18–22, 2024). Ed. by Weisong Shi, Deepak Ganesan, and Nicholas D. Lane. MobiCom ’24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 2248–2250. DOI: [10.1145/3636534.3695905](https://doi.org/10.1145/3636534.3695905).
- [27] Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*. Gurobi Optimization, LLC. 2025. URL: <https://www.gurobi.com>.
- [28] Sofonias Hailu, Mikko Säily, and Olav Tirkkonen. “RRC State Handling for 5G.” In: *IEEE Communications Magazine* 57.1 (Jan. 2019), pp. 106–113. DOI: [10.1109/MCOM.2018.1700957](https://doi.org/10.1109/MCOM.2018.1700957).
- [29] Ahmad Hassan, Wei Ye, Anlan Zhang, Jason Carpenter, Ruiyang Zhu, Shuowei Jin, Feng Qian, Z. Morley Mao, and Zhi-Li Zhang. “The Case for Boosting Mobile Application QoE via Smart Band Switching in 5G/xG Networks.” In: *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications* (San Diego, CA, USA, Feb. 28–29, 2024). Ed. by Nigel Davies and Chenren Xu. HotMobile ’24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 127–132. DOI: [10.1145/3638550.3641132](https://doi.org/10.1145/3638550.3641132).
- [30] Ahmad Hassan et al. “OPCM: Opportunistic Performance-driven Connectivity Management for 5G/xG Networks.” In: *Proc. ACM Netw.* 3.CoNEXT4 (2025), pp. 1–23. DOI: [10.1145/3768970](https://doi.org/10.1145/3768970).
- [31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Series in Statistics. New York: Springer, 2009. ISBN: 978-0-387-84858-7. DOI: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7).
- [32] Matthew J. Hausknecht and Peter Stone. “Deep Recurrent Q-Learning for Partially Observable MDPs.” In: *2015 AAAI Fall Symposia* (Arlington, VA, USA, Nov. 12–14, 2015). AAAI Press, 2015, pp. 29–37. URL: <https://cdn.aaai.org/ocs/11673/11673-51288-1-PB.pdf>.

- [33] Marvin Illian, Ramin Khalili, Antonio A. de A. Rocha, and Lin Wang. “Cells on Autopilot: Adaptive Cell (Re)Selection via Reinforcement Learning.” In: *2026 24th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. 🏆 **Best Paper Award**. Columbus, OH, USA: IEEE, June 2026, pp. 1–8. DOI: [10.23919/WiOpt71098.2026.11568224](https://doi.org/10.23919/WiOpt71098.2026.11568224).
- [34] Marvin Illian, Ramin Khalili, Antonio A. de A. Rocha, and Lin Wang. *Cells on Autopilot: Adaptive Cell (Re)Selection via Reinforcement Learning*. 2026. arXiv: [2601.04083](https://arxiv.org/abs/2601.04083) [cs.NI]. URL: <https://arxiv.org/abs/2601.04083>.
- [35] Marvin Illian, Ramin Khalili, Antonio A. de A. Rocha, and Lin Wang. “Cooperative Multi-Agent Reinforcement Learning for Idle-Mode Cell (Re)Selection.” In: *Proceedings of the 51st IEEE Conference on Local Computer Networks (LCN)*. IEEE LCN 2026. Coimbra, Portugal: IEEE, Oct. 2026, pp. 1–13. DOI: [10.1109/LCN67947.2026.11660770](https://doi.org/10.1109/LCN67947.2026.11660770).
- [36] Marvin Illian, Ramin Khalili, Antonio A. de A. Rocha, and Lin Wang. “Learning-Based Cell (Re)Selection for Cellular Networks.” ⓘ **Manuscript in preparation**. 2026.¹
- [37] Marvin Illian, Björn Luchterhandt, and Lin Wang. “Band Switching for Mobile Energy Optimization in 5G Networks and Beyond.” In: *Proceedings of the 20th Workshop on Mobility in the Evolving Internet Architecture*. MobiArch ’25. Hong Kong, China: Association for Computing Machinery, 2025, pp. 44–50. ISBN: 979-8-4007-1974-5. DOI: [10.1145/3737897.3767294](https://doi.org/10.1145/3737897.3767294).
- [38] Keyang Jiang, Qiang Wang, Yahao Xu, and Hongbin Deng. “Observation-Time-Action Deep Stacking Strategy: Solving Partial Observability Problems with Visual Input.” In: *2024 International Joint Conference on Neural Networks (IJCNN)*. Yokohama, Japan: IEEE, 2024, pp. 4932–4939. DOI: [10.1109/IJCNN60899.2024.10650736](https://doi.org/10.1109/IJCNN60899.2024.10650736).
- [39] Mads Lauridsen, Daniela Laselva, Frank Frederiksen, and Jorma Kaikkonen. “5G New Radio User Equipment Power Modeling and Potential Energy Savings.” In: *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*. ISSN:

¹ Manuscript to be submitted to IEEE Transactions on Networking. This journal article follows the fast-track publication opportunity associated with the WiOpt 2026 Best Paper Award.

- 2577-2465. Piscataway, NJ, USA: IEEE, Sept. 2019, pp. 1–6. DOI: [10.1109/VTCFall.2019.8891215](https://doi.org/10.1109/VTCFall.2019.8891215).
- [40] Mads Lauridsen, Laurent Noël, Troels Bundgaard Sørensen, and Preben Mogensen. “An Empirical LTE Smartphone Power Model with a View to Energy Efficiency Evolution.” In: *Intel Technology Journal* 18.1 (Mar. 2014), pp. 172–193. ISSN: 1535-864X. URL: https://vbn.aau.dk/ws/files/176790997/An_Empirical_LTE_Smartphone_Power_Model_with_a_View_to_Energy_Efficiency_Evolution.pdf.
- [41] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization.” In: *International Conference on Learning Representations* (New Orleans, LA, USA). OpenReview.net, 2019. URL: <https://openreview.net/forum?id=Bkg6RiCqY7> (visited on 06/30/2026).
- [42] Björn Luchterhandt. “Enhancing Cellular Handovers: Optimizing Quality of Experience using Machine Learning.” Faculty of Computer Science, Electrical Engineering and Mathematics; Computer Networks Group. Advised by Marvin Illian; submitted to Prof. Dr. Lin Wang and Dr. Simon Oberthür. Master’s thesis. Paderborn, Germany: Paderborn University, Feb. 2025. URL: <https://github.com/bjoluc/5g-handover-optimization/blob/main/thesis.pdf> (visited on 06/30/2026).
- [43] Mehrzad Malmirchegini, Mutaz Shukair, Peter Rached, Salil Sawhney, Mouaffac Ambriss, Kausik Ray Chaudhuri, and Sandip Sarkar. “Layer Management Through Idle-Mode Parameter Optimization in Multi-Carrier LTE Networks.” In: *2016 IEEE Wireless Communications and Networking Conference (WCNC)*. Doha, Qatar: IEEE, 2016, pp. 1–6. DOI: [10.1109/WCNC.2016.7564898](https://doi.org/10.1109/WCNC.2016.7564898).
- [44] Hongzi Mao, Malte Schwarzkopf, Shaileshh Bojja Venkatakrishnan, Zili Meng, and Mohammad Alizadeh. “Learning Scheduling Algorithms for Data Processing Clusters.” In: *Proceedings of the ACM Special Interest Group on Data Communication* (Beijing, China, Aug. 19–23, 2019). SIGCOMM ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 270–288. ISBN: 978-1-4503-5956-6. DOI: [10.1145/3341302.3342080](https://doi.org/10.1145/3341302.3342080).

- [45] Hongzi Mao, Shaileshh Bojja Venkatakrisnan, Malte Schwarzkopf, and Mohammad Alizadeh. *Variance Reduction for Reinforcement Learning in Input-Driven Environments*. 2018. arXiv: [1807.02264](#).
- [46] Ekaterina V. Markova, Varvara E. Manaeva, Elena Zhabankova, Dmitri Moltchanov, Pavel Balabanov, Yevgeni Koucheryavy, and Yuliya Gaidamaka. "Performance-Utilization Trade-Offs for State Update Services in 5G NR Systems." In: *IEEE Access* 12 (2024), pp. 129789–129803. DOI: [10.1109/ACCESS.2024.3442825](#).
- [47] Faris B. Mismar, Ahmad AlAmmouri, Ahmed Alkhateeb, Jeffrey G. Andrews, and Brian L. Evans. "Deep Learning Predictive Band Switching in Wireless Networks." In: *IEEE Trans. Wirel. Commun.* 20.1 (2021), pp. 96–109. DOI: [10.1109/TWC.2020.3023397](#).
- [48] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. "Asynchronous Methods for Deep Reinforcement Learning." In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. Vol. 48. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, June 2016, pp. 1928–1937. URL: <https://proceedings.mlr.press/v48/mniha16.pdf>.
- [49] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. *Playing Atari with Deep Reinforcement Learning*. 2013. arXiv: [1312.5602](#).
- [50] Michael S. Mollel, Shubi Kaijage, and Michael Kisangiri. "Deep Reinforcement Learning based Handover Management for Millimeter Wave Communication." In: *International Journal of Advanced Computer Science and Applications* 12.2 (2021), pp. 784–791. DOI: [10.14569/IJACSA.2021.0120298](#).
- [51] Prasad Nandkumar. *LTE RAN Ericsson-Huawei-Nokia Parameter Mapping*. No publication date. URL: <https://pdfcoffee.com/lte-ran-ericsson-huawei-nokia-parameter-mapping-byprasadnandkumar-pdf-free.html> (visited on 06/30/2026).

- [52] NGMN Alliance. *NGMN Recommendation on Self-Organising Networks (SON) and Operations & Maintenance (O&M) Requirements*. Requirement Specification. Next Generation Mobile Networks Alliance, 2008. URL: https://www.ngmn.org/wp-content/uploads/NGMN_Recommendation_on_SON_and_O_M_Requirements.pdf (visited on 06/30/2026).
- [53] NTIA. FCC, *NTIA Establish Spectrum Coordination Initiative*. 2022. URL: <https://www.ntia.gov/press-release/2022/fcc-ntia-establish-spectrum-coordination-initiative> (visited on 06/30/2026).
- [54] OpenStreetMap contributors. *OpenStreetMap data retrieved via Overpass API*. Data available under the Open Database License (ODbL). 2025. URL: <https://overpass-api.de/api/interpreter> (visited on 06/30/2026).
- [55] Rahul Arun Paropkari, Anurag Thantharate, and Cory Beard. "Deep-Mobility: A Deep Learning Approach for an Efficient and Reliable 5G Handover." In: *2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*. Chennai, India: IEEE, Mar. 2022, pp. 244–250. DOI: [10.1109/WiSPNET54241.2022.9767158](https://doi.org/10.1109/WiSPNET54241.2022.9767158).
- [56] Arunselvan Ramaswamy. "Gradient Clipping in Deep Learning: A Dynamical Systems Perspective." In: *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. Vol. 1. Lisbon, Portugal: SciTePress, Feb. 2023, pp. 107–114. ISBN: 978-989-758-626-2. DOI: [10.5220/0011678000003411](https://doi.org/10.5220/0011678000003411).
- [57] Stefan Schneider, Holger Karl, Ramin Khalili, and Artur Hecker. "Multi-Agent Deep Reinforcement Learning for Coordinated Multipoint in Mobile Networks." In: *IEEE Transactions on Network and Service Management* 21.1 (2024), pp. 908–924. ISSN: 1932-4537. DOI: [10.1109/TNSM.2023.3300962](https://doi.org/10.1109/TNSM.2023.3300962).
- [58] Stefan Schneider, Stefan Werner, Ramin Khalili, Artur Hecker, and Holger Karl. "mobile-env: An Open Platform for Reinforcement Learning in Wireless Mobile Networks." In: *IEEE/IFIP Network Operations and Management Symposium (NOMS)* (Budapest, Hungary, Apr. 25–29, 2022). IEEE, Apr. 2022, pp. 1–3. DOI: [10.1109/NOMS54207.2022.9789886](https://doi.org/10.1109/NOMS54207.2022.9789886).

- [59] Andrew Silva and Matthew C. Gombolay. “Encoding Human Domain Knowledge to Warm Start Reinforcement Learning.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 6. 2021, pp. 5042–5050. DOI: [10.1609/aaai.v35i6.16638](https://doi.org/10.1609/aaai.v35i6.16638).
- [60] Sunil. *Cell Reselection Procedure*. Techplayon. Mar. 9, 2024. URL: <https://www.techplayon.com/cell-reselection/> (visited on 06/30/2026).
- [61] Richard S. Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning. Cambridge, Massachusetts; London, England: The MIT Press, 2018. 526 pp. ISBN: 978-0-262-03924-6. URL: <https://mitpress.mit.edu/9780262039246/reinforcement-learning/>.
- [62] Muhammad Tayyab, George P. Koudouridis, Xavier Gelabert, and Riku Jäntti. “Signaling Overhead and Power Consumption during Handover in LTE.” In: *2019 IEEE Wireless Communications and Networking Conference (WCNC)* (Marrakesh, Morocco, Apr. 15–18, 2019). Piscataway, NJ, USA: IEEE, 2019, pp. 1–6. DOI: [10.1109/WCNC.2019.8885459](https://doi.org/10.1109/WCNC.2019.8885459).
- [63] Sethu Vijayakumar. *The Bias–Variance Tradeoff*. Lecture Notes. Original source no longer available; archived copy used. University of Edinburgh, 2007. URL: <https://web.archive.org/web/20131005010514/http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf> (visited on 06/30/2026).
- [64] Ronald J. Williams. “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning.” In: *Machine Learning* 8.3–4 (May 1992), pp. 229–256. DOI: [10.1007/BF00992696](https://doi.org/10.1007/BF00992696).
- [65] Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, and Huadong Ma. “Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption.” In: *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*. SIGCOMM ’20. Virtual Event, USA: Association for Computing Machinery, Aug. 2020, pp. 479–494. ISBN: 978-1-4503-7955-7. DOI: [10.1145/3387514.3405882](https://doi.org/10.1145/3387514.3405882).

- [66] Vijaya Yajnanarayana, Henrik Rydén, and László Hévizi. “5G Handover using Reinforcement Learning.” In: *3rd IEEE 5G World Forum (5GWF)* (Bangalore, India, Sept. 10–12, 2020). Piscataway, NJ, USA: IEEE, 2020, pp. 349–354. DOI: [10.1109/5GWF49715.2020.9221072](https://doi.org/10.1109/5GWF49715.2020.9221072).
- [67] Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. “Towards Understanding Convergence and Generalization of AdamW.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.9 (Sept. 2024), pp. 6486–6493. DOI: [10.1109/TPAMI.2024.3382294](https://doi.org/10.1109/TPAMI.2024.3382294).

COLOPHON

This document was typeset using the typographical look-and-feel classicthesis developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". classicthesis is available for both L^AT_EX and L^X:

<https://bitbucket.org/amiede/classicthesis/>

Happy users of classicthesis usually send a real postcard to the author, a collection of postcards received so far is featured here:

<http://postcards.miede.de/>

Thank you very much for your feedback and contribution.

Final Version as of September 8, 2026 (classicthesis v4.6).