Tampere University

Sharada Prabhashwara Jayaweera

NETWORK ENERGY SAVING IN 5G-ADVANCED AND BEYOND

Faculty of Information Technology and Communication Sciences (ITC)
Master's thesis
December 2025

Abstract

Sharada Prabhashwara Jayaweera: Network Energy Saving in 5G-Advanced and Beyond
Master's thesis
Tampere University
Master's Degree Programme in Computing Sciences and Electrical Engineering
April 2025

Energy consumption in 5G networks has emerged as a critical concern, with Base Stations (BS) accounting for a substantial portion of overall network power usage. While various energy-saving techniques have been proposed and studied, cell Discontinuous Transmission and Reception (DTX/DRX) optimization remains relatively under-explored despite its potential for significant Network Energy Saving (NES). This thesis addresses this gap by developing a Deep Reinforcement Learning (DRL) framework to dynamically optimize DTX/DRX ON duration across multiple cells.

A comprehensive System Level Simulator (SLS) was utilized and enhanced with multi-cell capabilities, interference modelling and a dynamic power consumption model. The DRL framework employs a single-agent approach with centralized coordination, enabling comprehensive network visibility and coordinated decision-making across cells. The reward function was carefully formulated to balance NES with Quality of Service (QoS) requirements.

Critical insights emerged during agent training, including reward misalignment issues where the agent optimized reward values rather than actual network performance. These challenges were addressed through reward function refinements and hyperparameter tuning. The results demonstrate that the proposed framework achieves considerable energy savings while maintaining QoS through dynamic DTX/DRX parameter optimization. Across diverse deployment configurations ranging from high-traffic to low-traffic scenarios, static power consumption was reduced by more than 50% compared to the Always ON baseline, whilst mean throughput degradation remained below 10%. Furthermore, delay characteristics remained acceptable across all configurations, confirming the viability of the DRL framework.

Keywords: 5G, Network Energy Saving, Cell Discontinuous Transmission and Reception, Deep Reinforcement Learning, Quality of Service, Energy Efficiency

Use of AI in Thesis

I have utilised AI tools in my thesis:

☐ No
☒ Yes

The AI tools utilised in my thesis, and their purposes are described below:

Names and versions of AI tools: ChatGPT (GPT-5), Claude (Sonnet 4.5) and Perplexity AI (Pro)

Purpose of using AI tools:

ChatGPT and Claude were used to brainstorm ideas, assist with coding, paraphrasing, text summarization and language polishing. Perplexity AI was used as a search engine for finding relevant literature and summarizing the content of those. All AI-generated content were critically reviewed and edited to reflect my ideas.

Sections where AI tools were used:

Throughout the whole thesis, AI was used for aforementioned purposes.

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

Preface

This thesis was completed as a commissioned project in collaboration with Intel Labs (Santa Clara, CA, US). I wish to express my sincere gratitude to my supervisors and examiners, Professor Sergey Andreev and Doctoral Researcher Yekaterina Sadovaya, for this valuable opportunity, their guidance, constructive feedback, and steady encouragement throughout the course of this work. Their academic insight and thoughtful advice were essential in defining the research direction and refining the final results.

I would also like to thank my colleagues at Intel Labs for their collaboration and support during this project. In particular, I am grateful to Dr. Shu-ping Yeh and Dr. Hosein Nikopour for their practical guidance, valuable discussions, and feedback that strengthened the applied aspects of this research.

I am also grateful to Tampere University for providing the facilities, technical resources, and an encouraging academic environment that made this research possible.

Finally, I extend my gratitude to my family and friends for their unwavering support and encouragement throughout this journey. I am also thankful to everyone who contributed to the completion of this thesis, either directly or indirectly. Their support and assistance have been of great importance throughout this process.

Sharada Prabhashwara Jayaweera
Tampere, Finland
December 2025

Contents

List of Symbols and Abbreviations

3GPP	The 3rd Generation Partnership Project
AI	Artificial Intelligence
AI/ML	Artificial Intelligence/Machine Learning
AR	Augmented Reality
ASM	Advanced Sleep Mode
BS	Base Stations
CDF	Cumulative Distribution Fucntion
DC-1	Deployment Configuration 1
DC-2	Deployment Configuration 2
DC-3	Deployment Configuration 3
DC-4	Deployment Configuration 4
DC-5	Deployment Configuration 5
DCI	Downlink Control Information
DEE	Data Energy Efficiency
DL	Downlink
DRL	Deep Reinforcement Learning
DTX/DRX	Discontinuous Transmission and Reception
DU	Data Unit
EC	Energy Consumption
EE	Energy Efficiency
EIRP	Effective Isotropic Radiated Power
eMBB	Enhanced Mobile Broadband
ETSI	European Telecommunications Standards Institute
FDD	Frequency Division Duplexing

FR1	Frequency Range 1
FR2	Frequency Range 2
ICI	Inter-Cell Interference
ICT	Information and Communication Technology
IEA	International Energy Agency
IoT	Internet of Things
ITU	International Telecommunication Union
KPI	Key Performance Indicator
LEE	Latency Energy Efficiency
MAC	Medium Access Control
ML	Machine Learning
mMIMO	Massive Multiple Input Multiple Output
mMTC	Massive Machine-Type Communications
mmWave	Millimeter-Wave
MN	Mobile Network
NES	Network Energy Saving
NG-RAN	Next Generation Radio Access Network (5G)
non-PHY	Non-Physical Layer
O-DU	Open Data Unit
O-RAN	Open Radio Access Network
O-RU	Open Radio Unit
OFDM	Orthogonal Frequency-Division Multiplexing
OPEX	Operating Costs
PA	Power Amplifier
PF	Proportional Fair

PHY	Physical Layer
PPO	Proximal Policy Optimization
PRB	Physical Resource Block
PSD	Power Spectral Density
QoS	Quality of Service
RAN	Radio Access Network
RF	Radio Frequency
RL	Reinforcement Learning
RRC	Radio Resource Control
RSRP	Reference Signal Received Power
RU	Radio Unit
RX	Receive
SCS	Sub-carrier Spacing
SLS	System Level Simulator
SM1	Sleep Mode 1 (Micro Sleep)
SM2	Sleep Mode 2 (Light Sleep)
SM3	Sleep Mode 3 (Deep Sleep)
SSB	Synchronization Signal Block
TDD	Time Division Duplexing
TRP	Transmission-Reception Points
TRX	Transceiver
TTI	Transmission Time Interval
TTL	Time-To-Live
TX	Transmit
UE	User Equipment

UL	Uplink
UMa	Urban Macro
URLLC	Ultra-Reliable Low Latency Communication
VR	Virtual Reality
WB	World Bank
Xn	Network Interface between NG-RAN Nodes

1 Introduction

1.1 Network Energy Saving in 5G/5G+ Systems

Global energy consumption continues to climb at a steady rate, driven by economic growth, industrial activity, and rapid urbanization. The International Energy Agency (IEA) reported a 2% rise in global energy demand in 2023 (it is expected to rise by the same rate in 2024) [1], pushing carbon emissions from energy production to a record high of 37.4 gigatons in 2023 [2]. Amidst this backdrop, industries are intensifying efforts to improve energy efficiency, with the Information and Communication Technology (ICT) sector emerging as a critical area of focus. The International Telecommunication Union (ITU) reports that about two-thirds of the world's population is now connected to the Internet [3]. While studies on the ICT sector's contribution to global carbon emissions vary, they generally estimate it to range from 1.5 to 4%. According to the data in the joint ITU and World Bank (WB) report, at least 1.7% of global emissions can be attributed to the ICT industry [4]. to put this figure in perspective, the aviation industry accounts for 2.5% of global emissions [5]. Despite accounting for approximately 1.7% of global greenhouse gas emissions currently, the ICT sector's contribution is increasing at a significant rate due to the surge in data traffic and the proliferation of connected devices [6].

The Ericsson Mobility Report (2024) predicts that global mobile data traffic will grow by a factor of 3 between 2024 and 2030, reaching a staggering 473 exabytes per month by 2030 [6]. This explosion in traffic is fueled by several factors, including the growing adoption of high-bandwidth applications such as high quality video streaming, Augmented Reality (AR), Virtual Reality (VR), autonomous vehicles, and cloud gaming. Additionally, the expansion of Internet of Things (IoT) ecosystems, with billions of sensors and devices being deployed for industrial automation, smart cities, and healthcare, adds to the burden on communication networks. As these trends continue, the deployment of 5G and Beyond technologies has become essential to meet the unprecedented demands for connectivity, low latency, and high data rates. By the end of 2024, 5G was projected to account for approximately 34% of global mobile data traffic, up from 25% at the close of 2023. This proportion is expected to rise significantly, reaching 80% by 2030 [6]. The deployment of 5G and Beyond networks represents a paradigm shift in communication technologies, enabling diverse use cases such as Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communication (URLLC), and Massive Machine-Type Communications (mMTC).

However, these advancements come with increased energy demands due to dense base station deployments, advanced signal processing requirements, and continuous operations to support dynamic traffic patterns.

In established markets, energy expenses account for up to 15% of Network Operating Costs (OPEX), while in developing markets, this percentage typically ranges from 15% to 30%. Traditional mobile networks present significant opportunities for enhancing energy efficiency. Only 15% of the energy consumed by these networks is allocated to forwarding data, meaning 85% is spent on activities that do not directly generate revenue [7]. The Radio Access Network (RAN) accounts for the largest share of energy consumption in cellular networks, contributing 80% of the total energy footprint [7]. The RAN consists of several components, including BS, antennas, and user devices. BS alone are responsible for the majority of the RAN's energy usage, as they remain active 24/7 to manage fluctuating traffic. However, they often remain idle for extended periods during a 24-hour cycle, consuming energy for tasks such as system broadcasts, idle resource maintenance like cooling, heating, lighting, and powering uninterruptible and auxiliary power supplies. Antennas, particularly Massive Multiple Input Multiple Output (mMIMO) configurations, require significant energy for beamforming and multi-user support. User devices also contribute to the overall consumption, with their energy requirements driven by factors such as signal strength, application demands, and mobility patterns.

The energy challenges posed by 5G networks have spurred significant research and standardization efforts. The 3rd Generation Partnership Project (3GPP) has consistently integrated energy-saving features into its recent releases. These efforts have encompassed a range of approaches, including lean carrier designs and Advanced Sleep Modes (ASMs) to minimize energy consumption during periods of low activity [8]. Dynamic spectrum sharing has been explored as a means to improve resource utilization and potentially reduce energy waste. Optimizations related to mMIMO beamforming and efficient management of dense network deployments have also been a focus. More recently, the use of Artificial Intelligence (AI) or Machine Learning (ML)-based techniques for network management has emerged as a key area of development, enabling more dynamic and intelligent control of energy consumption across various network functions. These techniques allow for finer-grained adjustments to network resources based on real-time traffic patterns and user demands. This ongoing commitment to energy efficiency has continued into the era of 5G Advanced, with a strong emphasis on minimizing energy consumption in the RAN [9].

1.2 Research Objectives and Contribution

Research on energy efficiency optimization has primarily focused on single-domain strategies targeting specific components of the network. For instance, studies like [10] focus on ASMs and strategies to implement it for optimized performance, while [11] proposes a Q-learning based approach to manage ASMs. Similarly, techniques like antenna muting have been used to deactivate idle antenna elements in mMIMO systems [12]. While these efforts have made substantial strides in optimizing individual components, they often address specific network aspects in relative isolation, highlighting the need for more holistic, system-wide energy management solutions.

Beyond component-specific optimization, RL applications in network optimization face several fundamental challenges, including multi-agent coordination complexity, and the difficulty of designing reward functions that balance energy efficiency with QoS requirements. Furthermore, cell DTX/DRX configuration optimization remains relatively under-explored compared to other ASM techniques. This work addresses these gaps by developing a single-agent RL framework that optimizes DTX/DRX parameters across multiple cells through centralized coordination. The approach aims to achieve network-wide energy savings while maintaining QoS.

The main contributions of this thesis are:

- Identifying the gaps, challenges and opportunities in NES on 5G networks.
- Adopting practical assumptions on the network dynamics, physical layer and quantifying their impact on the NES on 5G networks.
- Addressing the optimization of DTX/DRX ON durations via DRL-based solution using PPO algorithm, where optimized ON durations determine OFF period lengths that govern achievable sleep state depths in ASM progression, achieving static power reductions while maintaining near-baseline QoS across diverse traffic conditions.
- Providing recommendations on the solution's limitations and performance under different scenarios.

This document is organized into five chapters. Chapter 2 provides a detailed review of the related literature, 3GPP energy efficiency enablers, power consumption models, and metrics and Key Performance Indicators (KPIs) used in the study.

Chapter 3 introduces the simulator tool, the machine learning model, and the simulation scenarios. Chapter 4 presents an in-depth analysis of the results. Finally, Chapter 5 highlights the key contributions, summarizes the main findings and explores directions for future research.

2 Related Literature

This chapter reviews existing literature on NES in 5G-Advanced systems. The review begins with simulation-based and ML techniques for network optimization, followed by an identification of research gaps and challenges. Subsequently, 3GPP energy efficiency enabling features are examined, particularly ASMs, DTX/DRX mechanisms and AI/ML techniques. Finally, power consumption models, EE metrics and KPIs are discussed to establish the evaluation framework for this research.

2.1 Simulation-based Techniques

SLSs have become an indispensable tool for evaluating and optimizing 5G networks. Recent research demonstrates the widespread adoption of simulation-based methodologies for assessing network performance, resource allocation strategies, and QoS under varying deployment conditions. [13] has developed physical layer abstraction models on top of open source SLSs to enable scalable evaluation of large-scale 5G deployments. A flexible simulation framework named ”WiSE” has been introduced in [14] to support the analysis of diverse application scenarios including mMTC and URLLC. Calibration and validation of simulators based on 3GPP specifications is detailed in [15]. Furthermore, simulations have been employed to assess network behaviour in specialized contexts such as industrial automation environments with stringent QoS requirements [16], network slicing with dynamic resource allocation [17], interference management in shared spectrum scenarios [18], and mobility modeling in complex urban environments [7]. Similarly, SLSs have become fundamental for evaluating energy saving mechanisms in 5G networks. These tools replicate large-scale network operations by modeling traffic dynamics, user mobility, and resource management, enabling detailed exploration of the trade-offs between energy consumption and network performance under varied load conditions [8]. Nokia' white paper [19] demonstrates utilizing simulations to assess advanced sleep modes and antenna adaptation, achieving power savings while sustaining throughput and latency in realistic deployments. With recent advances in use of AI methodologies for network optimization, simulators have almost become a necessary component of network research. Studies such as [11], [20], [21] and [22] utilizes SLSs to train and evaluate their proposed approaches, underscoring the importance of simulation platforms capable of representing system-level interactions across both physical and higher-layer domains.

2.2 Machine Learning-based Techniques

The use of AI/ML techniques have become increasingly popular for networking related aspects. ML represents a set of algorithmic techniques that use data to build intelligent systems. The field consists of three main approaches: supervised learning, unsupervised learning, and reinforcement learning [8]. Large datasets have enabled supervised and unsupervised learning methods to analyze network traffic patterns and forecast 5G network behaviour [8]. RL operates differently from these approaches by focusing on learning decision-making strategies through ongoing interaction with the environment [8]. Consequently, ML techniques serve as essential enablers for intelligent mobile networks capable of environmental characterization, spatio-temporal prediction of system dynamics, and adaptive real-time responses [8].

Recent advances demonstrate that AI/ML methodologies enable automated network management, dynamic resource allocation, and self-organizing network capabilities that adapt to varying operational conditions [23][24]. Furthermore, these AI/ML techniques have been applied across diverse optimization domains including network slicing orchestration, real-time performance enhancement, fault detection, and end-to-end network efficiency improvements [25][26]. ML techniques have progressively advanced from supervised learning to RL and DRL for optimizing dynamic BS operations. RL techniques have proven particularly effective for NES due to their ability to learn adaptive policies through environmental interaction. DRL has been successfully employed for dynamic cell activation and deactivation strategies in O-RAN architectures, demonstrating substantial improvements in RAN EE [27]. RL-based approaches have been applied to centralized self-organizing network frameworks for energy management in heterogeneous cloud RANs [28], power control optimization in device-to-device communications [29], and energy-aware resource allocation for network slicing [30]. DRL frameworks such as [20] achieve up to 45% reduction in BS energy consumption by learning optimal DTX/DRX configurations under latency constraints. Industry trials by Vodafone and Ericsson confirm up to 33% network energy savings leveraging AI-driven management in realistic 5G environments [31]. The 3GPP *TR 338.743* outlines AI/ML enablers for NG-RAN targeting energy efficiency improvements, with attention to training algorithms and deployment architectures [32]. The Q-learning-based sleep mode management presented in [11] indicates practical benefits in reducing power demand under real traffic conditions . Although RL demonstrates significant potential for NES, the formulation of RL frameworks presents considerable difficulty. This complexity arises from multiple challenges, particularly the determination of appropriate reward functions, state spaces, and action spaces to ensure QoS performance while saving energy.

2.3 Gaps and Challenges

Several fundamental challenges characterize RL applications in network optimization. Environments are often partially observable, meaning agents lack complete system state information and can only access partial observations. In wireless networks, partial observability commonly occurs when a base station remains unaware of neighbouring base station loads. Multi-agent systems introduce additional complexity, where perceived rewards depend on actions taken independently by multiple agents [8]. Designing effective RL reward functions presents an ongoing challenge, as achieving appropriate trade-offs between energy consumption, latency, and other QoS metrics while preventing convergence to local optima proves difficult. While numerous studies have investigated direct ASM optimization, cell DTX/DRX configuration optimization remains relatively under-explored in the literature [20]. These gaps highlight the need for holistic frameworks integrating realistic simulations, and advanced RL frameworks suited for future 5G-Advanced systems.

This work addresses the above gaps through a single-agent framework that optimizes DTX/DRX parameters across multiple cells. The agent simultaneously controls DTX/DRX ON duration settings for all cells, emulating centralized multi-cell coordination. This approach provides the agent with comprehensive network visibility, enabling coordinated actions across cells to optimize overall network performance while achieving energy savings without degrading QoS.

.

2.4 3GPP Energy Efficiency Enabling Features

Improving energy efficiency has emerged as a critical driver for the evolution of mobile networks, especially as 5G-Advanced and research toward 6G intensify the focus on reducing both operational costs and environmental impact. The RAN is responsible for the majority of network power consumption, as highlighted previously, and recent efforts in the 3GPP standardization process, particularly in Release 18 and Release 19, reflect a strategic shift toward empowering networks to operate in a more dynamic, intelligent, and context-aware manner. These releases introduce a range of features that enable networks to adapt their energy usage as opposed to relying solely on static power allocation approaches that often lead to significant wastage of energy, especially during periods of low traffic.

Release 18 introduces a variety of techniques that can generally be divided into four domains: time domain, frequency domain, spatial domain, and power domain. [33]. The purpose of these techniques is to reduce the energy consumption of 5G mMIMO BS while maintaining the QoS for end users (throughput and latency) [19][34]. These techniques dynamically align base station operation with real-time network load and user demands, allowing deeper and more granular adaptation than previous releases. Although these techniques operate in different domains, they achieve energy savings through two main methods [19].

- **Enabling more sleep opportunities for cells** through advanced scheduling and reduced transmission activity during low traffic, such as DTX/DRX, and SSB-less secondary cell operation.
- **Enabling energy-efficient data transmission**, especially through adaptive power control and antenna muting in mMIMO, dynamically adjusting resource allocation based on actual traffic and QoS requirements.

The above techniques are primarily intended for situations with light to moderate user demand, where there is considerable potential to reduce energy consumption. In these conditions, the network infrastructure is often underutilized and full operational capacity across all cells is not necessary to maintain coverage. Even during peak usage, only about 30% of radio resources are actively used, while many cells experience long periods of low activity, especially early in the morning [19]. Mobile traffic is also distributed unevenly: roughly 30% of BS are responsible for carrying about 80% of the total traffic, while the other 70% handle only a small fraction, resulting in low utilization and thus offering greater opportunities for energy savings [19].

2.4.1 Advanced Sleep Modes (ASM)

In modern mobile networks, energy consumption remains considerable even when user activity is minimal, since BS continue to operate core functions regardless of traffic load. To mitigate this inefficiency, the concept of ASMs) has been introduced.

ASMs rely on the gradual deactivation of base station components, with the depth of sleep being determined by the time required for the shutdown and reactivation. This progressive approach enables a balance between energy savings and service responsiveness. Unlike LTE, where always-on reference signals limited the ability to suspend operations, the more flexible signal design of 5G allows BS to intermittently deactivate and only resume activity when necessary for synchronization, system broadcasts, or random access procedures [35].

[36] introduced four distinct sleep levels known as ASMs. These sleep modes are nicely summarized in [37] as included below.

- **SM1:** Operates at the level of an Orthogonal Frequency-Division Multiplexing (OFDM) symbol (71 s). The power amplifier and selected processing units are temporarily deactivated, enabling very fast transitions.
- **SM2:** Works on the subframe or Transmission Time Interval (TTI) scale (1 ms). More components enter the sleep state compared to SM1, achieving greater power reduction.
- **SM3:** Defined at the frame duration (10 ms). Most base station components are switched off, offering deeper energy savings with longer wake-up times.
- **SM4:** Applies at the scale of 1 s. The base station is largely inactive but maintains backhaul connectivity to allow reactivation when required.

Building on this foundation, 3GPP later formalized the concept by recognizing three sleep modes: **micro sleep, light sleep, and deep sleep** as outlined in [34]. In this thesis, these modes will be referred to as **SM1, SM2, and SM3** respectively to ensure clarity and uniformity in the subsequent discussion.

- **Micro Sleep (SM1):** In this state, Downlink (DL) transmissions and Uplink (UL) receptions are temporarily suspended. For analytical purposes in energy efficiency studies, transitions to and from the active state are considered instantaneous. Only a limited set of RF components are disabled, making this mode comparable to a stand-by condition [34][33].
- **Light Sleep (SM2):** In light sleep, DL transmissions and UL receptions are suspended, and the sleep period must be longer than the total transition time required to enter and exit the state. A large share of physical layer units, including the analog front-end and digital baseband blocks, are powered down [34][33].

- **Deep Sleep (SM3):** This mode halts DL transmission and UL reception for extended periods, but requires a longer inactivity period since the total transition time is higher. Almost all physical layer modules are turned off, with only essential elements such as the clock generator kept operational to preserve wake-up capability [34][33].

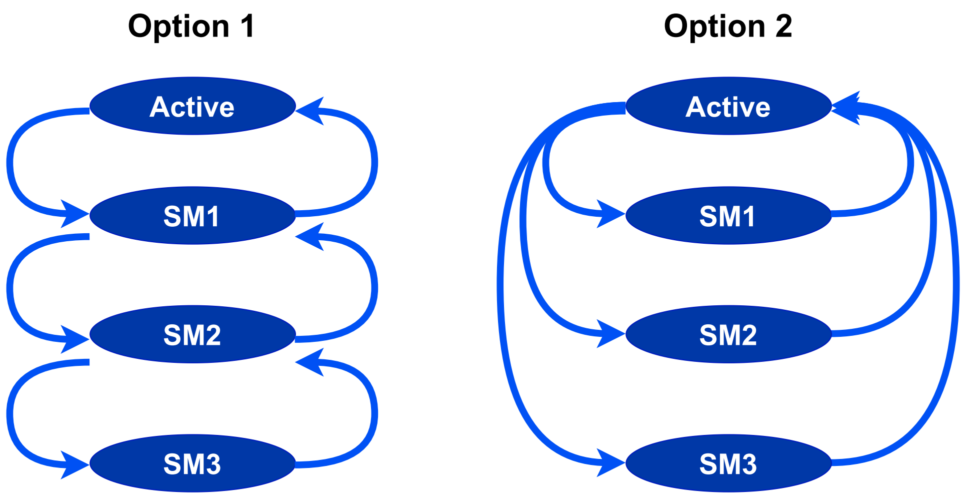

Figure 2.1 *Transition options between the sleep modes and active state. (Adapted from [33])*

Different transition strategies can be applied to sleep modes. [33] introduces two such options as illustrated in Figure 2.1. In *Option 1*, the base station progresses through the sleep modes sequentially, while in *Option 2*, it can switch directly from the active state to any sleep mode. In this work, the *Option 1* is adopted.

2.4.2 Cell Discontinuous Transmission and Reception (DTX/-DRX)

In 3GPP Release 18, the Cell DTX/DRX framework was introduced as a significant advancement for improving network energy efficiency. The principle behind this mechanism is to alternate cell operation between active and inactive phases in a repeating cycle as illustrated in Figure 2.2. During the active phase, the base station functions normally with full transmission and reception, while in the inactive phase, both data transfer and associated L1 channels are suspended, enabling the use of ASMs [20].

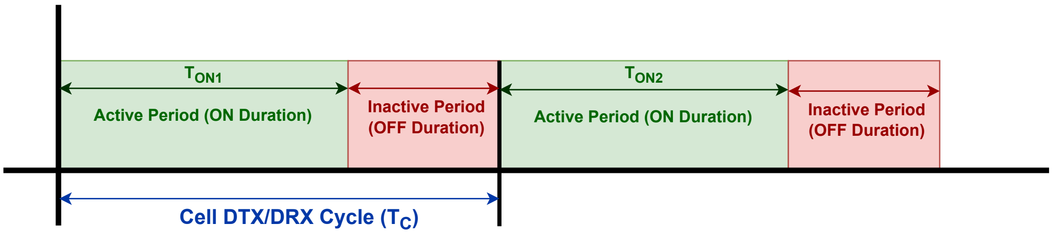

Figure *2.2 Operation of Cell DTX/DRX.*

Unlike full cell switch-off, cell DTX/DRX works on relatively fine time scales—ranging from the symbol level up to several hundred milliseconds [38], making it particularly useful in scenarios where completely turning off a cell is not feasible, as it still provides time-domain energy savings. Its operation is governed through 3GPP-defined signalling, including Radio Resource Control (RRC) and L1 Downlink Control Information (DCI) [20]. Key RRC configuration parameters cover aspects such as cycle periodicity, activation status, on-duration, and offsets for both cycles and active slots. However, for the purposes of this thesis, only RRC signalling is considered and the offsets are ignored to keep the analysis tractable.

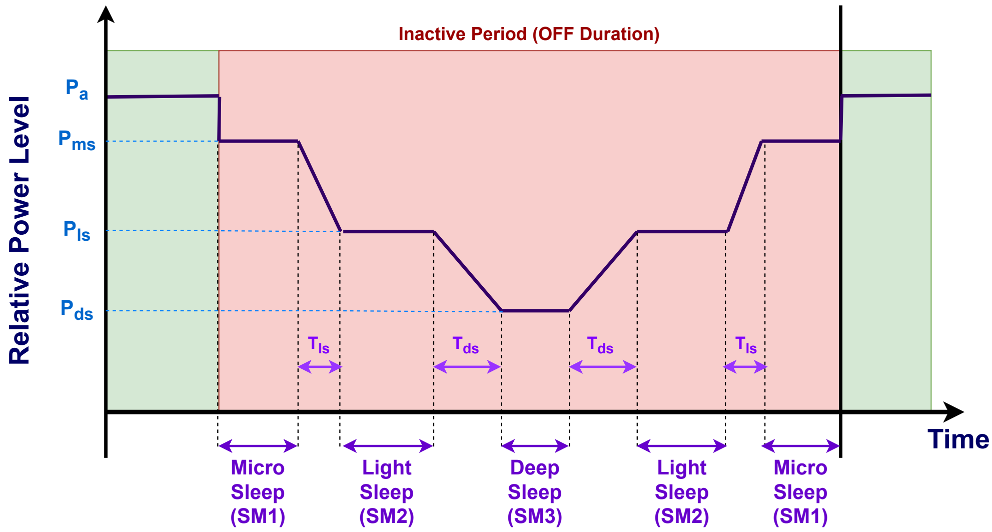

Figure *2.3 Sleep Modes Progression in DTX/DRX Inactive Period.*

During the OFF or the Inactive period of a DTX/DRX cycle, the base station may enter different sleep modes depending on the remaining duration of the cycle. Deeper sleep states can only be reached when the OFF duration is sufficiently long, since the base station must be able to return to the active state before the next

DTX/DRX cycle begins. The progression of these states is illustrated in Figure 2.3. It should be noted that the transition latency increases with the depth of the sleep mode, although the time required to enter and exit a given mode remains symmetrical. Detailed discussion of the corresponding power levels and transition times is provided in Section 2.5 in which the power consumption models are introduced.

2.4.3 Artificial Intelligence/Machine Learning (AI/ML)

3GPP has progressively integrated AI/ML into its framework for NES. The initial foundation was set in Release 17 with *TR 37.817*, which defined how NG-RAN nodes could collect and share data including energy-related data to support AI/ML-driven decision making [39]. This study established the concept of exchanging performance and energy metrics across the network to enable data-driven optimization, including cell deactivation, traffic offloading, and load adaptation. Release 18 transformed these concepts into normative specifications as highlighted in [32]. A key introduction is the Energy Cost metric, which is defined as an index that reflects the energy usage of an NG-RAN node. This metric can be obtained on demand and exchanged between NG-RAN nodes via the Xn interface and standardized as a way for NG-RAN nodes to express their energy consumption and share this information with peers [32]. Release 19 extends these capabilities by broadening the scope of AI/ML use cases and explicitly formalizing energy efficiency as a service criterion. The study *TR 22.882* [40] explores how energy efficiency and even carbon emission metrics can be exposed as part of communication service offerings.

As noted in [19], the 3GPP Release 18 Network Energy Saving framework offers significant potential for enhancement through AI/ML integration. AI/ML techniques can improve energy efficiency by anticipating factors such as user mobility and variations in cell load, which are critical for determining the optimal power state of a base station [19]. Furthermore, [41] suggests extending Release 18 NES discussion to incorporate beam-level AI/ML actions, allowing predictive switching of beams rather than entire cells, alongside forecasting preferred DTX/DRX patterns to identify precise sleep opportunities. These concepts point towards a more adaptive and fine-grained approach to energy management, providing a strong foundation for the DRL-based method presented in this thesis. The Chapter 3 details this approach, focusing on optimizing DTX/DRX active(or ON) durations to achieve enhanced energy efficiency while maintaining network performance.

2.5 Power Consumption Models

In order to analyze and evaluate energy saving strategies in 5G-Advanced and beyond, it is necessary to work with a standardized framework that captures how BS consume power under different conditions. Without such a framework, the results across the studies would not be comparable, since deployment configurations and traffic profiles vary widely. To address this, 3GPP *TR 38.864* [34] has defined reference configurations and operating states for BS, together with corresponding parameters for relative power, transition time, and transition energy. This section carries a discussion on these reference configurations and parameters.

	Set 1 (FR1)	**Set 2 (FR1)**	**Set 3 (FR2)**
Duplexing Method	TDD	FDD	TDD
System Bandwidth	100 MHz	20 MHz	100 MHz
Sub-carrier Spacing (SCS)	30 kHz	15 kHz	120 kHz
Number of Transmission-Reception Points (TRP)	1	1	1
Total Number of DL Transmit Radio Units (TX RUs)	64	32	2
Total DL Power Level	55 dBm	49 dBm	33 dBm
Total Number of UL Receive Radio Units (RX RUs)	64	32	2

Table 2.1 *Reference Configurations. (Adapted from [34])*

The 3GPP *TR 38.864* [34] introduces reference configurations that serve as the foundation of the BS power consumption model. These configurations, shown in Table 2.1, represent three deployment sets covering both Frequency Range 1 (FR1) and Frequency Range 2 (FR2). FR1 represents sub-6 GHz operation, and FR1 covers the mmWave bands from 24.25 GHz to 71 GHz. FR1 deployments correspond to wide-area coverage and lower frequency operation, while FR2 offers very large bandwidth and capacity but shorter propagation distances.The table specifies key system parameters such as the Duplexing Method, System Bandwidth, Sub-carrier Spacing (SCS), the number of Transmission-Reception Points (TRP), the number of DL Transmit (TX) Radio Units (RUs), the corresponding DL power levels, and the number of UL Receive (RX) RUs. For instance, Set 1 FR1 assumes a Time Division Duplexing (TDD) system with 100 MHz bandwidth and 30 kHz SCS, equipped with

64 TX and 64 RX RUs, delivering a total DL power of 55 dBm. By contrast, Set 2 FR1 represents a Frequency Division Duplexing (FDD) macro-type configuration with narrower 20 MHz bandwidth and 15 kHz SCS, operating with 32 TX and RX RUs at a lower DL power of 49 dBm. Set 3 FR2 models a high-frequency mmWave system with 120 kHz SCS, two TX and RX units, and a total TX power of 33 dBm (Here it should be noted that the Effective Isotropic Radiated Power (EIRP) is limited to 63 dBm and scales with the number of TX units) [34].

Power State	**Relative Power *P***	**Additional Transition Energy *E***	**Total Transition Time *T***
Deep Sleep (SM3)	P1	E1	T1
Light Sleep (SM2)	P2	E2	T2
Micro Sleep (SM1)	P3	0	0
Active DL	P4	N.A.	
Active UL	P5		

Table 2.2 *Power States. (Adapted from [34])*

Following the reference configurations, the model introduces the notion of power states as shown in Table 2.2. These states describe how a base station consumes energy depending on whether it is transmitting, receiving, or idle. *Active DL* and *Active UL* correspond to continuous DL transmission and UL reception respectively, and are denoted by relative power values *P4* and *P5*. The three sleep modes, SM1, SM2 and SM3 are denoted by relative power values of *P3*, *P2* and *P1* respectively. In SM1, the station halts transmissions but can return to active state instantaneously, without any additional energy or time cost. SM2 and SM3 disable more hardware components, which requires non-zero transition times of *T2* and *T1* and additional transition energies *E2* and *E1* respectively when returning to activity. These modelled states reflect the practical behaviour of hardware components. The same notations will be used for these relative power levels and transition times in the subsequent discussions.

Table 2.3 quantifies the relative power values introduced in Table 2.1 for different reference configurations and BS categories. *Category 1* typically represents macro BS with higher power ratings and a larger number of radio chains, while *Category 2* corresponds to smaller, lower-power systems. The data illustrate clear scaling trends. For example, *Set 1* in *Category 1* active DL transmission consumes *280*

Power State	**BS Category 1**			**BS Category 2**		
	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
Deep Sleep (SM3)	1			1		
Light Sleep (SM2)	25			2.1		
Micro Sleep (SM1)	55	50	38	5.5	5	3
Active DL	280	200	152	32	26	17.6
Active UL	110	90	80	6.5	5.8	4.2

Table 2.3 *Relative Power Values. (Adapted from [34])*

relative units, while the same state in Category 2 requires only *32*. Sleep modes are represented by significantly lower values, with SM3 normalized to 1 unit in both categories. This normalization makes it possible to translate conceptual operating modes into numerical estimates, since the actual power in each state can be obtained by multiplying the relative factors with the absolute power levels. In the context of this thesis, the focus is placed on the values for ***BS Category 1, Set 2***, which correspond to a macro-type base station configuration. This choice aligns with the simulator employed in the study (detailed in Chapter 3), which adopts the Urban Macro (UMa) channel model, ensuring consistency between the power consumption assumptions and the propagation environment used for performance evaluation. In addition, the simulator implementation fixes the absolute power for SM3 at 1 dBm.

Power State	**BS Category 1**	**BS Category 2**
Deep Sleep (SM3)	50 ms	10 s
Light Sleep (SM2)	6 ms	640 ms
Micro Sleep (SM1)	0 ms	0 ms

Table 2.4 *Transition Times. (Adapted from [34])*

Table 2.4 specifies the total transition times required to to enter and exit the sleep modes. For Category 1 base stations, SM3 transition is modelled as *50 ms*, while SM2 requires 6 ms. For Category 2 systems, however, the figures increase significantly, with SM3 taking *10 s* and SM2 *640 ms*. These values emphasize that the feasibility of using deeper sleep modes depends on both the expected idle duration and the hardware category. Again in the context of this thesis, the values for ***BS Category 1*** are used. It should also be noted that, in this work, the assumption is that during any transition period the power consumption corresponds to that of the next sleep state.

The 3GPP model provides explicit formulations for BS power consumption during active operation. These equations distinguish between a static baseline component, which is independent of configuration scaling, and a dynamic component, which varies with system bandwidth, number of Transceiver (TRX) units, and transmission characteristics.

For DL operation, the total power consumption $\boldsymbol{P^{DL}}$ is expressed as the sum of static and dynamic parts:

$$P^{DL} = P^{DL}_{static} + P^{DL}_{dynamic}. \tag{2.1}$$

- $\boldsymbol{P^{DL}_{static}}$ represents the static portion of power consumed whenever the base station is in an active DL state. This part does not scale with bandwidth or antenna configuration.The baseline assumption sets this value equal to *P3*. An alternative option is also recognized, where *P3* is multiplied by a factor of 1.5 [34]:

$$P^{DL}_{static} = P_3. \tag{2.2}$$

$$P^{DL}_{static} = 1.5 \times P_3. \tag{2.3}$$

- $\boldsymbol{P^{DL}_{dynamic}}$ represents the dynamic portion of power consumed whenever the base station is in an active DL state. It scales with the proportion of active TRX units, the utilized bandwidth relative to the maximum bandwidth, and the ratio of power spectral density with respect to the reference configuration. $\boldsymbol{P^{DL}_{dynamic}}$ is defined as:

$$p^{DL}_{\text{dynamic}} = s_a \cdot \left(P_{\text{dyn,ante}} + \frac{s_f \cdot s_p}{\eta(s_f, s_p)} \cdot P_{\text{dyn,joint}} \right), \tag{2.4}$$

  where $\boldsymbol{s_a}$ is the proportion of active RUs, $\boldsymbol{s_f}$ is the ratio between utilized bandwidth and the maximum bandwidth, $\boldsymbol{s_p}$ is the ratio of PSD per TX RU between the DL transmission and the reference configuration. $\boldsymbol{\eta(s_f, s_p)}$ models the efficiency of the Power Amplifier (PA) under those conditions.

  $\boldsymbol{P_{dyn,ante}}$ denotes the portion of dynamic power that varies solely with the number of active antenna elements:

$$P_{\text{dyn,ante}} = A \cdot (P_4 - P_{static}). \tag{2.5}$$

$\boldsymbol{P_{dyn,joint}}$ represents the portion of dynamic power that scales jointly with the active antenna count, the occupied bandwidth, and the efficiency of the power amplifier:

$$P_{\text{dyn,joint}} = (1 - A) \cdot (P_4 - P_{static}) \,. \tag{2.6}$$

The parameter $\boldsymbol{A}$ is introduced as a split factor between the two dynamic power components. Its baseline setting is *0.4*, while alternative values of *0.1* and *0.7* are also suggested for evaluation in order to capture different behaviours.

In the baseline case, the PA efficiency factor $\boldsymbol{\eta(s_f, s_p)}$ is taken as 1 for all combinations of $\boldsymbol{s_a}$ and $\boldsymbol{s_p}$. For evaluation purposes, however, an alternative setting is introduced to reflect the effect of PA and resource utilization. In this case:

$$\eta(s_f, s_p) = \begin{cases} 0.76, & \text{if } s_f \cdot s_p \leq 0.5, \\ 1, & \text{otherwise.} \end{cases} \tag{2.7}$$

For UL operation, the total power consumption $\boldsymbol{P^{UL}}$ is expressed as the sum of static and dynamic parts similar to the DL case:

$$P^{UL} = P^{UL}_{static} + P^{UL}_{dynamic}. \tag{2.8}$$

- $\boldsymbol{P^{UL}_{static}}$ represents the static portion of power consumed whenever the base station is in an active UL state. This part does not scale with reference configurations.The baseline assumption sets this value equal to *P3*:

$$P^{UL}_{static} = P_3. \tag{2.9}$$

- $\boldsymbol{P^{UL}_{dynamic}}$ represents the dynamic portion of power consumed whenever the base station is in an active UL state. It scales with the reference configurations. $\boldsymbol{P^{UL}_{dynamic}}$ is defined as:

$$p^{UL}_{\text{dynamic}} = s_a \cdot P^{UL}_{\text{dyn,ante}}, \tag{2.10}$$

where $\boldsymbol{s_a}$ is the proportion of active RUs. $\boldsymbol{P^{UL}_{dyn,ante}}$ denotes the portion of dynamic power that varies with the number of active antenna elements:

$$P^{UL}_{\text{dyn,ante}} = P_5 - P_3. \tag{2.11}$$

It should be noted that the dynamic components are bounded. For the DL case, the following condition applies:

$$P_{\textbf{dynamic}}^{\textbf{DL}} \leq P_4 - P_{\textbf{static}}^{\textbf{DL}}. \tag{2.12}$$

Similarly, for the UL case the bound is given by:

$$P_{\textbf{dynamic}}^{\textbf{UL}} \leq P_5 - P_{\textbf{static}}^{\textbf{UL}}. \tag{2.13}$$

The power consumption model presented in 3GPP *TR 38.864* should be viewed as a reference framework rather than an exhaustive description of all possible base station behaviours. Its simplified assumptions make it suitable for systematic evaluation, while acknowledging that real deployments may vary in the number of supported states, their specific power characteristics, and the time required for transitions [34]. In this thesis, these standardized elements are adopted as the foundation for simulation and analysis, ensuring that subsequent evaluation of energy saving strategies remains both analytically manageable and sufficiently representative of real BS operation.

2.6 Energy Efficiency Metrics and KPIs

Energy Efficiency (EE) has become a central dimension in the evaluation of 5G-Advanced and beyond networks, and several standardization bodies have addressed the need to define measurable indicators. Among the most widely cited are the specifications of ETSI, 3GPP, and the O-RAN Alliance, each of which provides a different level of granularity. ETSI *ES 203 228* [42] focuses on energy efficiency assessment in mobile networks with an emphasis on the radio access domain. 3GPP *TS 28.5544* [43] incorporates energy efficiency as one of the seven standardized key performance indicators for end-to-end 5G systems. The O-RAN specification on Network Energy Savings Procedures and Performance Metrics [44] provides detailed metrics for RUs and Data Units (DUs) when energy saving functions are applied. These establish a layered view of how energy efficiency can be quantified, from component level to network slice and system wide analysis.

The ETSI specification [42] outlines a set of approaches for evaluating the energy efficiency of mobile network equipment and deployments. At the center of this framework is the concept of Energy Consumption (EC). EC is defined as the total energy drawn by all components of the mobile network, aggregated over the chosen observation interval and expressed in joules. Building on this, the specification introduces three principal performance metrics: *Data Volume, Coverage Area, and Latency.* Data Volume corresponds to the quantity of user data delivered by the network equipment under assessment during the chosen observation interval, expressed in bits. Coverage Area is considered a relevant metric for networks primarily designed to maximize geographical service reach, particularly in RU environments, and is measured in square meters. Latency is included as an additional performance indicator for networks in which URLLC use cases are predominant, and it is measured in milliseconds. EE is then formulated as the ratio between a given performance metric and the corresponding EC. In general form, this can be written as:

$$\text{Energy Efficiency (EE)} = \frac{\text{Performance Metric}}{\text{Energy Consumption (EC)}}. \tag{2.14}$$

Building on the above formulation from ETSI, the O-RAN Alliance specification [44] introduces a number of EE KPIs that focus on the behaviour of the Open Radio Unit (O-RU) and its interaction with the Open Data Unit (O-DU). These metrics are intended to capture how data throughput, latency, and radio transmission power relate to the actual energy consumed by the O-RU, thereby providing a fine-grained view of efficiency at the component level.

The most intuitive formulation is the *Data Energy Efficiency (DEE)*, which relates the user data volume carried over the network to the total energy consumption of the Mobile Network (MN):

$$\text{DEE} = \frac{\text{DV}_{\text{MN}}}{\text{EC}_{\text{MN}}}, \tag{2.15}$$

where $\boldsymbol{DV_{MN}}$ denotes the total UL and DL data volume, and $\boldsymbol{EC_{MN}}$ is the corresponding energy consumption of the MN.

Since O-RAN aims to characterize efficiency at the O-RU level, this expression is refined by considering the volume of Physical Resource Blocks (PRBs) processed by the O-RU. The resulting KPIs is referred to as *DEE for the O-RU*:

$$\text{DEE}_{\text{RU}} = \frac{\text{PRBV}_{\text{RU}}}{\text{EC}_{\text{RU}}}, \tag{2.16}$$

where $\boldsymbol{PRBV_{RU}}$ represents the number of PRBs sent or received by the O-RU in a given interval, and $\boldsymbol{EC_{RU}}$ is the energy consumption of the O-RU during that same interval.

Beyond throughput, latency also plays a significant role in efficiency assessments, particularly in URLLC communication scenarios. To account for this, the specification [44] defines the *Latency Energy Efficiency (LEE)*, which combines average end-to-end latency with total energy consumption:

$$\mathrm{LEE} = \frac{1}{T_{e2e} \times \mathrm{EC}_{\mathrm{MN}}}, \tag{2.17}$$

where $\boldsymbol{T_{e2e}}$ is the mean end-to-end latency experienced by user packets, and $\boldsymbol{EC_{MN}}$ is again the total network energy consumption. The reciprocal form ensures that lower latency and lower energy consumption both contribute positively to the metric.

An O-RU-specific variant of the eq. (2.17) KPI is defined as:

$$\mathrm{LEE}_{\mathrm{RU}} = \frac{1}{T_{e2e,\mathrm{RU}} \times \mathrm{EC}_{\mathrm{RU}}}, \tag{2.18}$$

where $\boldsymbol{T_{e2e,RU}}$ represents the average latency for user packets traversing the O-RU, while $\boldsymbol{EC_{RU}}$ captures the associated energy consumption of the O-RU.

Finally, the specification [44] introduces a RF based formulation that directly compares the total transmitted RF power with the energy consumed by the O-RU:

$$\mathrm{EE}_{\mathrm{DL}} = \frac{\text{Total RF power into antennas}}{\text{Total O}-\text{RU power consumption}}, \tag{2.19}$$

which reflects the efficiency of converting electrical input power into radiated output power.

Building on these standardized formulations, the remainder of this thesis employs a KPI conceptually similar to DEE. This KPI, shown in eq. (2.20), will serve as the basic optimization target in the DRL framework developed in Chapter 3.This ensures that the learned policies are directed toward improving energy efficiency. However, modifications and enhancements to this basic KPI is employed to ensure better performance and robustness. The detailed discussion on this is carried out in Chapter 3 and Chapter 4.

$$\text{Energy Efficiency(EE)} = \frac{\text{Total Data Volume(DV)}}{\text{Total Energy Consumption (EC)}}. \tag{2.20}$$

While this KPI forms the basis for agent training, additional metrics are required to provide a comprehensive evaluation of the trained policies. Three such indicators are considered. The first is the *Mean Perceived Throughput*, which reflects the data rate as experienced by end users. The second is the *Mean Delay*, measured on a per-packet basis, which captures the responsiveness of the network and its ability to support latency-sensitive services. The third is the *Mean (Static) Power Consumption of the BS*, provides a measure of the baseline energy cost associated with maintaining network operation. Together, these metrics complement the efficiency KPI by allowing the performance of the reinforcement learning agent to be assessed not only in terms of NES but also in relation to user experience.

3 Methodology

This chapter presents the methodology employed for developing and evaluating the DRL-based energy saving framework for 5G networks. It begins with an overview of the SLS enhancements, including the multi-cell network topology, interference modelling, and DTX/DRX implementation. The RL framework is then described, detailing the algorithm configuration, state space representation, action space design, and the reward function formulations. The chapter concludes with the experimental setup, specifying the deployment configurations, training procedures, and performance evaluation metrics used to assess the proposed approach.

3.1 Simulation Tool and Configurations

To investigate the performance and energy efficiency of 5G-Advanced and beyond networks, it is necessary to employ tools that can capture system-level interactions across both the physical and higher-layer domains. Real-world testbeds are often costly and limited in scope, while purely analytical models may oversimplify the complexity of modern cellular systems. For this reason, our research group has developed a SLS, which provides a flexible and standards-aligned platform for modeling 5G networks based on 3GPP specifications. The simulator enables the evaluation of a wide range of scenarios by combining realistic channel, protocol, and network models with configurable parameters.

The architecture of the SLS is shown in Figure 3.1. At a high level, it is organized into Physical Layer (PHY) and Non-Physical Layer (non-PHY) domains. The non-PHY modules represent network layer and MAC layer functions, such as packet generation, topology management, and scheduling. These allow traffic flows to be created between User Equipment (UE) and BS, while also determining how resources are allocated in time and frequency. On the PHY side, the simulator incorporates detailed models of the wireless channel, including fading, mobility effects, and link blockages, as specified in 3GPP *TR 38.801* [45]. Antenna characteristics and power control mechanisms are also modeled, ensuring that the impact of physical layer behavior on network performance is accurately reflected. Together, the PHY and non-PHY components allow the SLS to replicate the interplay between protocol-level decisions and radio-level constraints in a realistic manner.

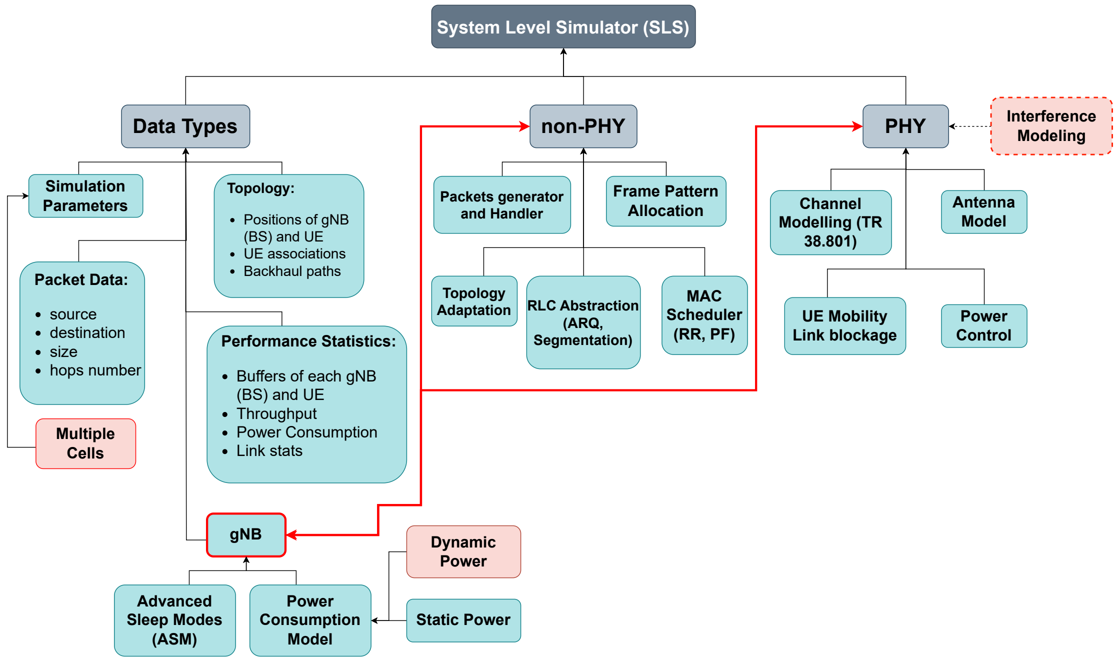

Figure 3.1 *System Level Simulator Block Diagram.*

In addition to modeling network operation, the simulator collects a set of performance statistics, such as throughput, delay, buffer occupancy, PRB utilization, and power consumption. These outputs are essential for evaluating trade-offs between efficiency and performance under different configurations. The framework is designed to be highly configurable, allowing the simulation of various topologies, packet-level traffic patterns, and cell deployments. Such flexibility is particularly important for studying energy-saving mechanisms like ASMs, where performance must be assessed across a range of operating conditions.

The baseline SLS or the parts of it has been used in the works done by our research group such as [46], [47] and [48]. As part of this research, several modifications were made to the baseline SLS developed by our research group. These changes were introduced to ensure that the simulator could accurately represent the energy behavior of 5G-Advanced and beyond networks and to provide a realistic environment for developing and testing the proposed DRL framework. Each enhancement was guided by a specific limitation identified in the earlier version of the simulator and by the practical requirements of evaluating energy efficient and robust network operation. The modules highlighted in pink color in Figure 3.1 correspond to these contributions.

The first enhancement was the inclusion of the dynamic component of the BS power consumption model, following the formulation defined in 3GPP TR 38.864 [34] and described in Section 2.5 of Chapter 2. The previous implementation considered only the static power component of the BS power consumption model, which meant that variations in load, bandwidth utilization, and antenna activity had no impact on the calculated energy usage. This simplification restricted the accuracy of relationship between total energy consumed and network load, which are central to this study. By introducing the dynamic term, the simulator now captures how energy consumption changes with operational conditions. This modification provides the DRL framework with a more meaningful reward signal that directly reflects the network's instantaneous energy state, allowing the learning agent to develop adaptive and context-aware control strategies. Refer to the set of equations eq. (2.1) to eq. (2.13) for more context.

The second enhancement extended the simulator's scope from single-cell to multi-cell operation. This change was motivated by the need to model the spatial and co-operative aspects of energy saving in practical network deployments. Real world networks are characterized by dense layouts and overlapping coverage, where the actions of one cell influence the load and energy state of its neighbours. Enabling multi-cell simulation makes it possible to investigate inter-cell coordination, load balancing, and cell activation or deactivation policies, which are key mechanisms for network-level energy optimization. In the revised implementation, the number of cells can be specified as a configuration parameter, with the option to assign either uniform or heterogeneous settings to each cell. This flexibility allows the impact of both homogeneous and mixed deployment conditions to be studied systematically.

The third addition was the introduction of interference modelling. This was an essential step toward improving the realism and robustness of the simulation framework. Inter-Cell Interference (ICI) plays a decisive role in determining both spectral efficiency and energy efficiency, particularly in dense 5G and beyond deployments. Without modelling interference, the simulator would tend to overestimate achievable gains from energy saving strategies, leading to optimistic but unrealistic results. The inclusion of interference modelling ensures that the DRL agent learns control policies that remain effective under realistic signal conditions.

Although traffic offloading is not examined within the scope of this thesis, the addition of multi-cell capability and interference modelling lays the groundwork for such studies in the future. The interference modelling framework is described in detail in the following section. With these features in place, the simulator can be extended to analyse user and traffic redistribution strategies across neighbouring cells, providing a basis for further research on adaptive load management and network-level energy optimization.

3.2 Interference Modelling

In realistic cellular network deployments, interference management plays a crucial role in determining overall system performance and resource utilization efficiency. The original SLS employed a static interference value for all UE across the network, which failed to capture the dynamic and spatially-varying nature of ICI present in actual 5G deployments. To address this limitation and enhance the fidelity of the simulation environment, a comprehensive interference modelling framework was implemented based on sectored antenna patterns and distance-dependent propagation characteristics.

The enhanced interference model introduces several configurable parameters that allow for flexible and realistic representation of ICI scenarios. Table 3.1 summarizes the key configurable parameters of the interference modelling framework. The model can be enabled or disabled based on simulation requirements, providing backward compatibility with legacy simulation configurations. A critical parameter is the antenna sector beam width, which is set to 120 degrees by default, representing the typical three-sector configuration commonly deployed in UMa cell scenarios. The number of sectors is automatically determined based on the configured sector beam width, ensuring consistency in the geometric representation. Additionally, specific sectors can be selectively enabled or disabled to model various deployment scenarios, though for the purpose of this thesis, all sectors are assumed to be active throughout the simulation period. Two threshold parameters govern the interference calculation: a maximum distance threshold that defines the spatial extent beyond which interference contributions are considered negligible, and an interference power threshold for categorizing strong interferers, set at 8 dBm based on the IEEE 802.16m Evaluation Methodology Document [49].

The implementation of the interference model follows a systematic approach that accounts for both geometric and propagation characteristics of the wireless channel. The computational procedure, outlined in Algorithm 1, begins by determining the distance and direction from each base station to every user equipment in the network based on their respective positions. Links exceeding the maximum distance threshold are immediately discarded to reduce computational complexity and focus on significant interference sources. For the remaining links, the azimuth angle is

Parameter	Description
Interference Modeling Enable/Disable	Toggle for interference model activation
Antenna Sector Beam Width	Angular width of each sector (default: 120°)
Number of Sectors	Determined by sector beam width
Enabled Sectors	Specific sectors to include in interference calculations
Maximum Distance Threshold	Spatial extent for interference consideration
Strong Interference Power Threshold	Threshold for significant interferers (8 dBm)

Table 3.1 *Configurable Parameters for Interference Modelling.*

calculated to determine which antenna sector the direction falls into, considering the sectorized radiation pattern of the base station antenna. Links are subsequently filtered based on whether their corresponding sectors are enabled in the current configuration. The received interference power is then calculated by accounting for path loss, antenna gain patterns, and shadowing effects. Finally, interferers below the strong interference power threshold of 8 dBm are classified as weak interferers and omitted from the total interference calculation, while the remaining contributions are aggregated to obtain the total interference power experienced by each UE.

The integration of the interference model with the DRL framework required careful consideration of the temporal dynamics introduced by the DTX/DRX cycles. As the DRL agent adjusts the active durations of different cells to optimize energy efficiency, the interference changes dynamically based on which cells are transmitting at any given time. To capture this relationship, an interference impact estimation mechanism was developed that maps the agent's actions to expected interference levels. The key assumption underlying this mechanism is that the interference from cell $\boldsymbol{j}$ to cell $\boldsymbol{k}$ is proportional to the duty cycle of cell $\boldsymbol{j}$, expressed as:

$$d_j = \frac{T_j^{ON}}{T_j^{C}}, \tag{3.1}$$

where $\boldsymbol{T_j^{ON}}$ represents the active duration and $\boldsymbol{T_j^{C}}$ represents the cycle duration. Consequently, the interference experienced by cell $\boldsymbol{k}$ due to cell $\boldsymbol{j}$ can be expressed as:

$$I_k = d_j \cdot I_{kj}, \tag{3.2}$$

where $\boldsymbol{I_{kj}}$ represents the baseline interference that cell $\boldsymbol{j}$ causes to cell $\boldsymbol{k}$ when continuously active.

Algorithm 1 Interference Calculation Procedure.

```
for all BS j and UE k do
    Calculate distance and direction from BS position to UE position
    if distance > maximum distance threshold then
        Discard link
        continue
    end if
    Calculate azimuth angle
    Determine sector based on azimuth angle
    if sector is not enabled then
        Discard link
        continue
    end if
    Calculate path loss, antenna gain, and shadowing
    Compute received interference power I_kj
    if I_kj < 8 dBm then
        Classify as weak interferer
    else
        Add to total interference power
    end if
end for
```

Interference levels between all cell pairs are precomputed assuming a duty cycle of 1 (Always ON scenario). These raw interference measurements are collected and stored as cell-specific statistics, forming a comprehensive interference profile for the network topology. From these statistics, Cumulative Distribution Fucntions (CDFs) are generated to characterize the statistical distribution of interference levels between each pair of cells. To provide a representative metric suitable for reinforcement learning observations, the 50th percentile value is extracted from each CDF, representing the median interference level between cells. Table 3.2 illustrates the structure of the resulting interference mapping table, where each entry $(\boldsymbol{i}, \boldsymbol{j})$ contains the median interference $\boldsymbol{I_{ij}}$ experienced by cell $\boldsymbol{i}$ due to cell $\boldsymbol{j}$. This mapping table remains static throughout the training process and serves as a lookup mechanism for estimating interference in the observation space.

From Cell / To Cell	Cell 1	Cell 2	Cell 3
Cell 1	-	I_{12}	I_{13}
Cell 2	I_{21}	-	I_{23}
Cell 3	I_{31}	I_{32}	-

Table 3.2 *Interference Mapping Table.*

In the reinforcement learning training loop, the interference estimation mechanism operates as follows. When the agent selects active duration actions for all cells in the network, the expected interference for cell $\boldsymbol{k}$ is calculated as:

$$I_k = \sum_{i \neq j} I_{kj}\, d_j, \tag{3.3}$$

which represents the weighted sum of interference contributions from all other cells, scaled by their respective duty cycles. This estimated interference value is incorporated into the observation space, providing the agent with awareness of the interference conditions resulting from its actions. However, it is important to note that during the actual simulation episodes, the precise interference is calculated using the full propagation model described in Algorithm 1, accounting for the actual positions of user equipment, current channel conditions, and instantaneous transmission states of all base stations. This dual approach of using estimated interference for observations while calculating actual interference for performance evaluation is aimed at ensuring the agent learns efficient policies based on realistic interference trends while maintaining simulation accuracy in the reward computation. The network performance metrics are consequently influenced by the actual calculated interference, which in turn affects the reward signal that guides the agent's learning process.

3.3 RL Framework

The optimization of DTX/DRX parameters in multi-cell 5G networks presents a complex decision-making problem characterized by high-dimensional state spaces, dynamic traffic patterns, and intricate trade-offs between EE and QoS. Traditional rule-based approaches often rely on fixed thresholds or heuristics that fail to adapt to varying network conditions and traffic loads. DRL offers a promising alternative by enabling an agent to learn optimal policies through interaction with the network environment. The proposed DRL framework follows the standard RL flow illustrated in Figure 3.2, where an agent interacts with the network environment by observing the current state, selecting actions based on its learned policy, and receiving rewards along with the subsequent state from the environment. Among the various DRL algorithms available, Proximal Policy Optimization (PPO) was selected as the primary learning algorithm for this work. PPO has demonstrated superior performance in continuous control tasks and offers several practical advantages, including stable training through clipped policy updates, efficient sample utilization through multiple epochs of minibatch updates, and robust convergence properties that make it well-suited for complex environments with high-dimensional observation spaces. Additionally, PPO's ability to handle both discrete and continuous action spaces provides flexibility in modelling the DTX/DRX configuration problem. The implementation leverages the Stable-Baselines3 library [50], which provides well-tested and optimized implementations of state-of-the-art RL algorithms.

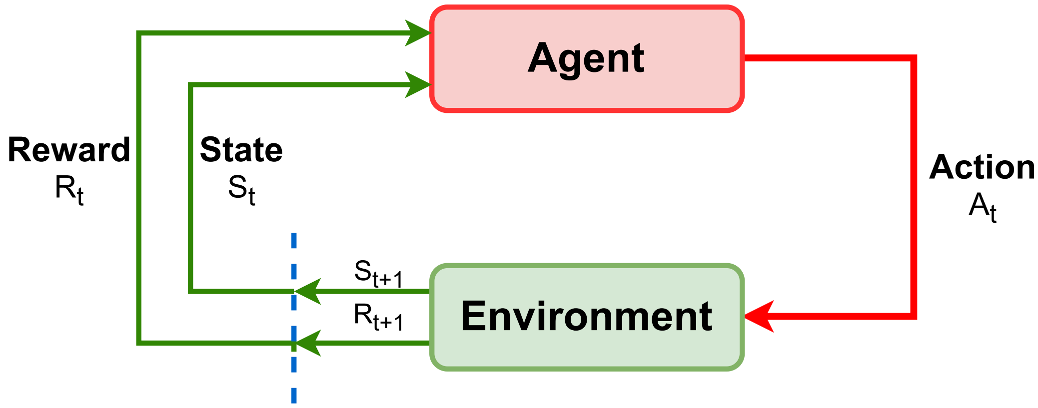

Figure* *3.2 Schematic Flow Diagram of RL.

3.3.1 Observation Space

The observation space captures the essential state information required for the agent to make informed decisions about DTX/DRX configurations across all cells in the network. The design philosophy emphasizes providing comprehensive yet concise representations of traffic conditions, resource utilization, power consumption, and interference levels. Table 3.3 summarizes the observation space, which is formulated as a dictionary space containing ten distinct features, each represented as arrays of dimension equal to the number of cells in the network.

The traffic-related observations include DL and UL buffer occupancy (`load_dl` and `load_ul`), which indicate the number of packets awaiting transmission in each cell's queues. These features provide the agent with awareness of pending traffic demands. Complementing the buffer occupancy, the burst throughput features (`tp_burst_dl` and `tp_burst_ul`) capture the instantaneous data rate achieved in the most recent transmission opportunity, bounded by a maximum throughput of 100 Mbps. These metrics enable the agent to assess both the volume of pending traffic and the rate at which it can be served.

Feature	Description	Dimension
`load_dl`	DL buffer occupancy	N_{cells}
`load_ul`	UL buffer occupancy	N_{cells}
`tp_dl`	DL perceived throughput	N_{cells}
`tp_ul`	UL perceived throughput	N_{cells}
`prb_util_dl`	DL PRB utilization	N_{cells}
`prb_util_ul`	UL PRB utilization	N_{cells}
`prev_tot_power_dl`	Previous DL transmit power	N_{cells}
`prev_tot_power_ul`	Previous UL transmit power	N_{cells}
`on_duration`	Current ON duration	N_{cells}
`interference_metric`	Normalized interference level	N_{cells}

Table 3.3 *Observation Space Components.*

The traffic-related observations include DL and UL buffer occupancy (`load_dl` and `load_ul`), which indicate the number of packets awaiting transmission in each cell's queues. These features provide the agent with awareness of pending traffic demands. Complementing the buffer occupancy, the perceived throughput features (`tp_dl` and `tp_ul`) capture the user-perceived data rate calculated based on the most recent transmission activity. The perceived throughput metric reflects the effective service rate experienced by users, accounting for both the volume of data delivered and the time required for transmission. These metrics enable the agent to assess both the volume of pending traffic and the QoS being delivered to users.

Resource utilization is characterized through the PRB utilization features (`prb_util_dl` and `prb_util_ul`), which represent the fraction of physical resource blocks actively used for data transmission in each direction. These normalized metrics, ranging from 0 to 1, provide insight into how effectively the available spectrum is being exploited. The power consumption features (`prev_tot_power_dl` and `prev_tot_power_ul`) record the total transmit power observed in the previous DTX/DRX cycle for each cell. These observations enable the agent to track the energy implications of its previous actions.

The current DTX/DRX configuration is reflected in the `on_duration` feature, which stores the active duration configured for each cell in seconds. This feature is bounded by the cycle duration, allowing values up to the maximum cycle period. Providing the agent with awareness of its previous actions facilitates temporal reasoning and enables learning of consistent policies.

A critical component of the observation space is the `interference_metric`, which quantifies the expected inter-cell interference level based on the current DTX/-

DRX configuration of all cells. This metric is computed using the interference mapping table described in Section 3.2, where the baseline interference between each cell pair is scaled by the duty cycle of the interfering cells. Mathematically, for cell $\boldsymbol{j}$, the interference metric is calculated as the ratio of the current expected interference to the maximum possible interference when all cells are continuously active. The normalization to the range $[\mathbf{0}, \mathbf{1}]$ is particularly important because raw interference power values are typically very small, and this scaling provides the RL framework with well-conditioned input features that facilitate effective learning.

3.3.2 Action Space

The action space defines the set of DTX/DRX configurations that the agent can select for each cell in the network. A key design decision in this work was the choice between continuous and discrete action representations. While continuous action spaces offer fine-grained control over the active durations, discrete action spaces were ultimately adopted for several compelling reasons. First, practical 5G systems operate with discrete ON duration values as specified in 3GPP *TS 38.331* [38]. The standard specifies a finite set of allowable ON duration values to ensure interoperability and simplified signalling. Second, discrete actions significantly improve training efficiency by reducing the exploration space and providing clear, unambiguous decisions. Third, discrete formulations align more closely with the actual operation of real network equipment, enhancing the practical applicability of the learned policies.

The discrete action space is formulated as a *MultiDiscrete* space, where each cell independently selects an ON duration from a predefined set of values. Table 3.4 presents the enumerated ON duration values, expressed in milliseconds, that constitute the action space. These values are taken from 3GPP *TS 38.331* [38], with one notable addition: a **320** ***ms*** duration value was included to enable the agent to select an ***"Always ON"*** configuration as the cycle duration for this study is set to **320** ***ms***. This augmentation ensures that the agent has the flexibility to completely disable sleep mode when traffic demands or other conditions calls for continuous transmission.

1	2	3	4	5	6	8	10	20	30	40	50	60	80
100	120	200	300	320	400	500	600	800	1000	1200	1600		

Table 3.4 *Discrete ON Duration Action Values (milliseconds).*

The action space dimensionality scales linearly with the number of cells, as each cell requires an independent ON duration decision. For a network with $\boldsymbol{N}$ cells and $\boldsymbol{K}$ possible ON duration values, the total number of unique joint actions is $\boldsymbol{K^N}$, representing all possible combinations of per-cell configurations. During each training step, which corresponds to the start of a new DTX/DRX cycle, the agent observes the current network state and outputs an action vector of length $\boldsymbol{N}$, where each element is an index into the set of allowable ON durations. This centralized decision-making approach enables the agent to coordinate DTX/DRX configurations across cells, potentially learning policies that account for inter-cell dependencies and interference relationships.

3.3.3 Reward Function

The design of the reward function is crucial in RL, as it encodes the optimization objectives and guides the agent's learning process toward desired behaviours. In the context of DTX/DRX optimization, the fundamental challenge lies in balancing multiple competing objectives: minimizing energy consumption to enhance network efficiency, maintaining QoS to ensure user satisfaction, and avoiding wasteful idle states where cells remain active without serving traffic. The reward function design evolved through an iterative development process, with each formulation addressing limitations discovered during preliminary training experiments. Three distinct reward formulations were developed and evaluated, leading to insights about the nuances of reward shaping for network energy optimization problems.

Reward Function 1: EE Ratio

The initial reward function adopted a straightforward ratio-based approach that directly captures the energy efficiency of the network as discussed in Section 3.2 eq. (2.20). The reward for cell $\boldsymbol{n}$ is computed as:

$$R_n = \frac{\text{Total\_Packets}_n/\text{Packets\_Max}}{\text{EC}_n}, \tag{3.4}$$

where $\boldsymbol{Total\_Packets_n}$ represents the number of packets successfully transmitted during the cycle, $\boldsymbol{Packets\_Max}$ provides normalization based on the maximum buffer size across all cells, and $\boldsymbol{EC_n}$ denotes the normalized energy consumption. The energy consumption is calculated by integrating the mean transmit power over the cycle duration and normalizing by the maximum possible energy consumption if the cell were to transmit at maximum power throughout the entire cycle.

An optional idle penalty term was included to discourage cells from remaining active when no traffic is present. The idle penalty for cell $\boldsymbol{n}$ is computed as:

$$\text{Penalty}_n = -\frac{\text{Idle\_Time}_n}{T_{c,n}}, \tag{3.5}$$

where $\boldsymbol{Idle\_Time_n}$ represents the duration during which cell $\boldsymbol{n}$ was active but had no packets to transmit, and $\boldsymbol{T_{c,n}}$ is the cycle duration for cell $\boldsymbol{n}$. This penalty is subtracted from the EE ratio in eq. (3.4). The complete per-cell reward including the penalty is:

$$R_n = \frac{\text{Total\_Packets}_n/\text{Packets\_Max}}{\text{EC}_n} - \frac{\text{Idle\_Time}_n}{T_{c,n}}. \tag{3.6}$$

The total reward provided to the agent is obtained by summing the per-cell rewards across all cells in the network:

$$R = \sum_{n=1}^{N} R_n. \tag{3.7}$$

During preliminary training experiments with this reward formulation, the agent exhibited erratic and seemingly random behaviour as presented in Chapter 4. Analysis of the training episodes revealed that the agent was not effectively learning to balance NES with QoS. As the reward function only considered successfully transmitted packets in the numerator, it does not provide direct visibility into packet drops. Packets that exceeded their delay budget and were discarded had no explicit representation in the reward signal. Consequently, the agent could inadvertently learn policies that aggressively minimized active durations to reduce energy consumption, leading to excessive packet drops due to insufficient transmission opportunities. Without explicit feedback about these dropped packets, the agent lacked the information necessary to understand the QoS implications of its actions.

Reward Function 2: EE Ratio with Dropout Awareness

To address the limitation identified in Reward Function 1, a modified formulation was developed that explicitly accounts for packet drops. The reward for each cell $\boldsymbol{n}$ is computed as:

$$R_n = \frac{\text{Total\_Packets}_n/(\text{total\_Packets}_n + \text{Dropped\_Packets}_n)}{\text{EC}_n}, \tag{3.8}$$

where $\boldsymbol{Total\_Packets_n}$ represents the number of packets successfully transmitted during the cycle, $\boldsymbol{Dropped\_Packets_n}$ accounts for packets that exceeded their delay budget and were discarded, and $\boldsymbol{EC_n}$ denotes the normalized energy consumption. This normalization approach ensures that the reward reflects the true packet delivery success rate relative to the attempted transmissions. By including dropped packets in the denominator, the formulation explicitly penalizes configurations that lead to excessive packet drops due to insufficient active durations. The objective was to force the agent to minimize dropped packets and thus ensure minimal or no blockage for UE. The energy consumption calculation and idle penalty follow the same approach as Reward Function 1.

While this modification is expected to improve the agent's awareness of QoS degradation, training experiments revealed a fundamental flaw in the ratio-based reward structure that persisted in both Reward Functions 1 and 2. The critical issue surfaces during periods when no traffic is present in the network. In such scenarios, both $\boldsymbol{Total\_Packets_n}$ and $\boldsymbol{Dropped\_Packets_n}$ equal zero, causing the numerator of the reward expression to become zero regardless of the denominator value. Consequently, the reward evaluates to zero whether the agent chooses to keep cells active (wasting energy) or puts them to sleep (conserving energy optimally). This creates a null gradient situation where the agent receives identical reward signals for dramatically different behaviours. Even with the idle penalty term included, the reward function fails to provide meaningful differentiation between sleep and wasteful active states during idle periods. This ambiguity led to random, sub-optimal policy learning.

Reward Function 3: Conditional Linear Formulation

The insights gained from the limitations of ratio-based formulations motivated the development of a fundamentally different reward structure. This was further motivated by linear reward formulation presented in [20]. Recognizing that optimal behaviour differs qualitatively depending on whether traffic is present or absent, the novel reward function employs a conditional architecture that adapts its optimization criteria to the backlog state. This design explicitly addresses the zero-numerator problem by using entirely different reward expressions for different traffic conditions.

When no backlog exists across the network ($\boldsymbol{B_{tot} = 0}$), where $\boldsymbol{B_{tot}}$ represents the total buffer occupancy summed across all cells and both UL and DL directions, indicating an idle period with no packets awaiting transmission, the reward directly encourages sleep states and penalizes unnecessary active periods:

$$R = w_{\text{sleep}} \cdot \text{Sleep\_sum} - w_{\text{idle}} \cdot \text{Idle\_sum}, \tag{3.9}$$

where $\boldsymbol{Sleep\_sum}$ represents the total fraction of time spent in sleep mode across all cells, $\boldsymbol{Idle\_sum}$ captures the total fraction of active time with no transmission activity across all cells, and $\boldsymbol{w_{sleep}}$ and $\boldsymbol{w_{idle}}$ are weighting parameters. In this formulation, sleeping provides a positive reward contribution proportional to the aggregate sleep fraction, while remaining active without purpose incurs a penalty. Note that in this conditional structure, the idle penalty is not a separate additive term but is inherently integrated into the reward design and the $\boldsymbol{Idle\_sum}$ term directly penalizes wasteful active time. This structure provides clear, non-zero gradients that guide the agent toward energy-efficient behaviour during idle periods.

Conversely, when backlog is present ($B_{\text{tot}} > 0$), indicating that packets are awaiting transmission in at least one cell, the reward shifts focus to service delivery and energy efficiency:

$$R = w_{\text{service}} \cdot \frac{T_{\text{tot}}}{T_{\text{max}}} - w_{\text{energy}} \cdot \frac{E_{\text{tot}}}{E_{\text{max}}} - w_{\text{idle}} \cdot \text{Idle\_sum}, \tag{3.10}$$

where $\boldsymbol{T_{tot}}$ denotes the total bits successfully delivered across all cells in both UL and DL, $\boldsymbol{E_{tot}}$ represents the total energy consumed across all cells, and $\boldsymbol{T_{max}}$ and $\boldsymbol{E_{max}}$ are running normalizers that adapt to maintain stable reward scales throughout training. The idle penalty term ($\boldsymbol{w_{idle} \cdot Idle\_sum}$) is explicitly included in this expression as well, penalizing cells that remain active but do not actually transmit data. This addresses the scenario where traffic exists in the network but certain cells may have empty buffers yet remain unnecessarily active. The weighting parameters enable fine tuning of the relative importance of different objectives. Table 3.5 presents the reward weight configuration used in this work.

Parameter	**Value**
w_{sleep}	1.0
w_{idle}	0.5
w_{service}	1.0
w_{energy}	0.5

Table 3.5 *Reward Weight Parameters for Reward Function 3.*

The normalization terms $\boldsymbol{T_{max}}$ and $\boldsymbol{E_{max}}$ are dynamically updated during training using an exponential moving average mechanism. After each step, these normalizers are updated as:

$$\begin{aligned} T_{\max} &\leftarrow \max(0.99 \cdot T_{\max}, T_{\text{tot}} + \epsilon), \\ E_{\max} &\leftarrow \max(0.99 \cdot E_{\max}, E_{\text{tot}} + \epsilon), \end{aligned} \tag{3.11}$$

where $\boldsymbol{\epsilon}$ is a small constant to prevent division by zero. This adaptive scaling ensures that the reward components remain balanced even as the agent's policy evolves and the magnitude of achieved throughput and energy consumption changes. Unlike the ratio-based formulations, this structure maintains meaningful gradients across all traffic conditions, enabling the agent to learn consistent policies that appropriately balance service delivery and energy consumption.

The evolution from Reward Function 1 through Reward Function 3 illustrates the importance of careful reward shaping in RL for network optimization. Reward Function 1 provided a conceptually simple EE metric but lacked visibility into QoS degradation through packet drops. Reward Function 2 addressed the packet drop visibility issue but inherited the fundamental zero-numerator problem that prevents effective learning during idle periods. Reward Function 3 resolves both limitations through its conditional structure and additive formulation, providing clear learning signals across all network states.

3.3.4 Training Procedure

The RL agent is trained using the PPO algorithm with carefully selected hyperparameters to ensure stable and efficient learning. Table 3.6 summarizes the key training configuration parameters used in this work.

Parameter	**Value**
Policy Network	MultiInputPolicy
Learning Rate	3×10^{-4}
Entropy Coefficient	5×10^{-3}
Number of Steps per Update	1024
Episode Length (cycles)	128
Total Training Timesteps	1×10^{5}
Evaluation Frequency	512 steps

Table 3.6 *PPO Training Hyperparameters.*

An important aspect of the training framework is the modified reset mechanism employed in the custom Gymnasium environment. Unlike traditional RL environments in domains such as game playing, where the `reset()` function returns the system to a fixed initial state, the reset implementation in this work performs a soft reset that primarily clears statistical containers used to collect performance statistics while preserving the underlying network state and traffic patterns. This design choice reflects the continuous nature of network operation, where there is no meaningful ***"starting position"*** to return to. The episode length is configured to 128 DTX/DRX cycles, providing the agent with sufficient temporal context to learn policies that account for traffic dynamics over multiple cycles rather than making myopic decisions based on a single cycle observation.

During training, the agent's performance is periodically evaluated using a separate evaluation environment instance. The evaluation procedure operates in deterministic mode, where the agent selects actions based on the mean of the learned policy distribution rather than sampling stochastically. Evaluation episodes are executed every 512 training steps, and the best-performing model based on cumulative reward is saved for subsequent analysis. After training completion, the learned policy is deployed by loading the saved model weights and executing it in the simulation environment for a desired number of steps. At each step, the agent observes the current network state, predicts an action using the deterministic policy, and the environment advances by one DTX/DRX cycle based on the selected configuration.

The training process generates detailed logs of the agent's learning progress, including episodic returns, value function estimates, policy entropy, and various network performance metrics collected by the simulator. These logs enable analysis of the learning dynamics and facilitate identification of potential issues such as reward instability, policy collapse, or insufficient exploration. The integration of TensorBoard [51] logging provides real-time visualization of training metrics, supporting informed decisions about hyperparameter adjustments or training duration.

3.4 Considered Scenario and Parameters

The evaluation of the proposed DRL framework for DTX/DRX optimization requires careful configuration of the simulation environment to represent realistic 5G network conditions. This section describes the network deployment scenario, traffic configuration, and the various experimental configurations used for training the RL agent and assessing its performance under diverse operating conditions.

3.4.1 Network Deployment and Configuration

The simulation environment models an UMa cell deployment consisting of three hexagonal cells arranged in a cluster topology. Each cell has a radius of 1000 meters and is served by a BS equipped with a uniform planar array antenna system. The carrier frequency is set to 5 GHz with a total bandwidth of 100 MHz, representing a typical 5G mid-band deployment. UEs are distributed randomly within each cell's coverage area.

The antenna configuration at each base station consists of a 16×16 uniform planar array, providing a total of 256 antenna elements in both horizontal and vertical dimensions. While the simulator supports dynamic optimization of the number of active antenna elements (with configurations ranging from 2×2 to 16×16), the RL experiments maintain a fixed configuration of 16×16 active elements. This design choice allows the RL agent to focus exclusively on DTX/DRX ON duration optimization without the added complexity of joint antenna and power management optimization, which constitutes a separate optimization problem.

Resource allocation and scheduling decisions are managed by a Proportional Fair (PF) scheduler, which balances throughput maximization with fairness considerations among competing users. The DTX/DRX cycle duration is configured to 320 ms. This cycle duration also represents the periodicity at which the RL agent selects ON duration actions for each cell in the network.

3.4.2 Traffic Model and Load Configuration

Network traffic is generated using a model based on Poisson processes, where each UE independently generates traffic arrivals according to its own Poisson process. All UEs share the same arrival rate parameter λ, but their actual packet generation times differ due to independent random sampling. Each traffic burst consists of 10 packets and individual packets have a size of 1,053 bytes. Packets have a Time-To-Live (TTL) constraint of 20 ms, after which they are dropped if not successfully transmitted, representing delay-sensitive application requirements.

The traffic intensity in the network is controlled through two parameters: the number of active UE per cell and the Poisson arrival rate λ. Increasing λ causes each UE to generate traffic more frequently, while increasing the number of UEs adds more independent traffic sources to the network. The combination of these parameters determines the aggregate traffic load, which manifests as a specific level of PRB utilization. For instance, with 40 users per cell and $\lambda = 2$, the observed DL PRB utilization reaches approximately 80%, representing a moderately high load scenario. This relationship between user count, arrival rate, and PRB utilization enables systematic control of network load conditions for experimental purposes.

3.4.3 Training Configuration

The primary training configuration, Deployment Configuration 1 (DC-1), represents the baseline scenario under which the RL agent learns its DTX/DRX optimization policy. The key parameters for DC-1 are detailed in Table 3.7, with a visual representation of the deployment provided in Figure 3.3. This configuration is designed to expose the agent to realistic traffic conditions with high resource utilization, facilitating the learning of policies that must balance EE with QoS requirements. The configuration consists of 3 cells with 40 UEs per cell and a traffic intensity parameter $\lambda = 2$, resulting in an observed DL PRB utilization of approximately 80%. This utilization level represents a challenging scenario where the network experiences sustained traffic demand, requiring the agent to make tricky decisions about when to enter sleep states and when to maintain active transmission.

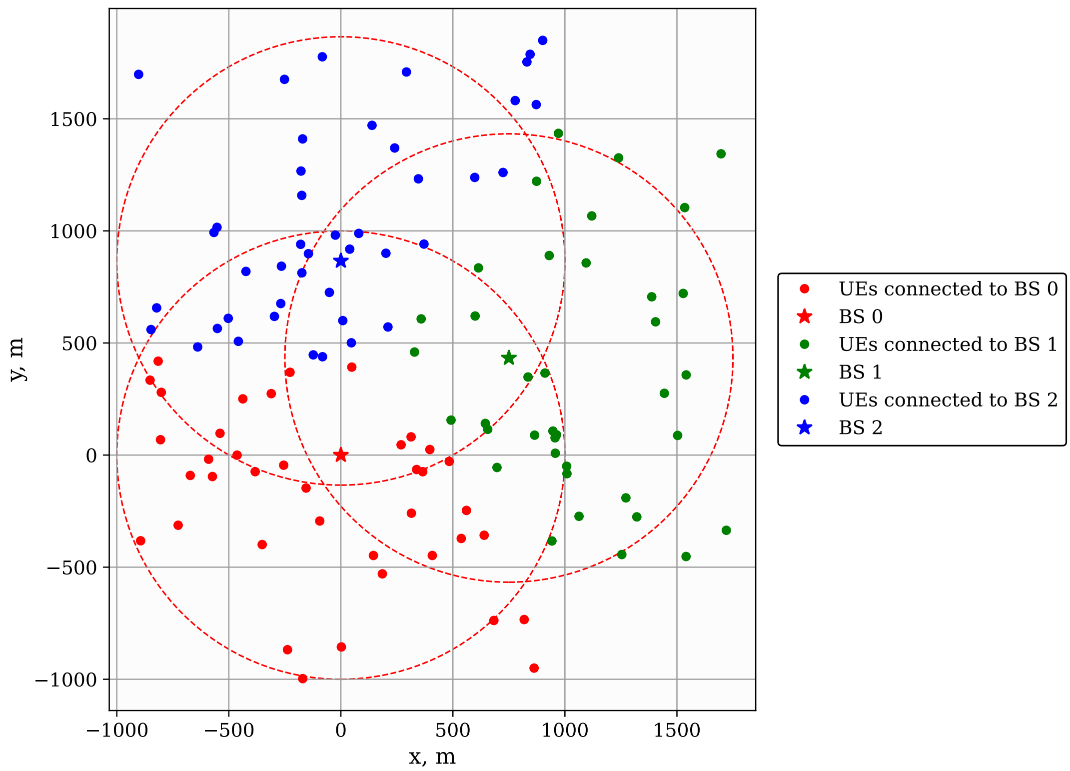

Figure 3.3 *Visual Representation of the Deployment Configuration.*

The agent undergoes training for 100,000 timesteps, where each timestep corresponds to one DTX/DRX cycle. Throughout training, the agent observes network state features, selects ON duration actions for all cells, and receives reward signals based on the reward functions described in Section 3.3.3. The training process employs the PPO algorithm with the hyperparameters specified in Section 3.3.4, and performance is periodically evaluated every 512 steps to track learning progress and identify the best-performing model.

Parameter	**Value**
Number of Cells	3
UEs per Cell	40
Traffic Intensity (λ)	2
Cell Radius	1000 m
Carrier Frequency	5 GHz
Bandwidth	100 MHz
Active Antenna Elements	16×16
Scheduler	Proportional Fair
DTX/DRX Cycle Duration	320 ms
Packet TTL	20 ms
Observed PRB Utilization	$\approx 80\%$
Total Training Timesteps	100,000
Random Seed	5

Table 3.7 *Training Configuration Parameters DC-1.*

3.4.4 Evaluation and Robustness Study

To assess the generalization capability and robustness of the learned policy, the trained agent is evaluated across multiple deployment configurations that differ in traffic intensity and network density. These evaluation scenarios, designated as DC-2 through DC-5, systematically vary the number of UEs and traffic arrival rates to create networks with substantially different load characteristics. Table 3.8 presents the complete set of evaluation configurations along with their observed PRB utilization levels.

Deployment Configuration 2 (DC-2) represents a medium traffic scenario with 25 UEs per cell at $\lambda = 2$, yielding approximately 50% PRB utilization. This configuration tests the agent's ability to achieve greater energy savings in a less congested network while maintaining QoS. Deployment Configuration 3 (DC-3) increases the load density with 50 UEs at $\lambda = 2$, resulting in approximately 90% PRB utilization.

Configuration	**Users per Cell**	λ	**PRB Util.**	**Description**
DC-1	40	2	$\approx 80\%$	Training/Baseline
DC-2	25	2	$\approx 50\%$	Medium Traffic
DC-3	50	2	$\approx 90\%$	Dense Traffic
DC-4	15	2	$\approx 30\%$	Sparse Traffic
DC-5	5	2	$\approx 10\%$	Very Sparse Traffic

Table 3.8 *Evaluation Deployment Configurations.*

This high-load scenario evaluates whether the agent can maintain QoS when the network approaches capacity, requiring longer or more frequent active periods.

Deployment Configuration 4 (DC-4) and Deployment Configuration 5 (DC-5) examine performance under sparse traffic conditions. DC-4 configures 15 UEs at $\lambda = 2$, achieving 30% PRB utilization, while DC-5 employs 5 UEs at $\lambda = 2$, resulting in only 10% PRB utilization. These low-load scenarios are particularly important for EE evaluation, as they present opportunities for substantial NES through aggressive sleep mode utilization. The agent must learn to recognize these idle periods and minimize unnecessary active time while remaining responsive to sporadic traffic arrivals.

The evaluation methodology compares the trained agent's performance across all five deployment configurations, analyzing key metrics including static power consumption, perceived throughput, packet delay, and PRB utilization. Detailed performance results and comparative analysis are presented in Chapter 4. This cross-configuration evaluation reveals the extent to which the policy learned under DC-1 conditions generalizes to substantially different operating conditions, providing insights into the robustness and adaptability of the learned DTX/DRX optimization strategy.

4 Results

This chapter presents a comprehensive evaluation of the proposed DRL framework for EE in multi-cell 5G networks. The chapter is organized into five sections. First, the behavior of the SLS is demonstrated to establish its alignment with theoretical expectations. Subsequently, the performance of the trained DRL agent is presented across different reward function formulations. The agent was trained using the PPO algorithm on a 3-cell network with deployment configuration DC-1 (40 UEs per cell, 80% PRB utilization). The trained policies are then evaluated across all deployment configurations (DC-1 through DC-5) to assess generalization capabilities. Performance is compared against an Always-ON baseline where the DTX/DRX ON duration equals the cycle duration, representing maximum capacity operation without energy saving. Three reward function formulations are examined, each representing different approaches to balancing EE and QoS objectives. The first two reward functions revealed fundamental limitations in their design, while the third reward function demonstrated improved performance. Additionally, a modified version of the third reward function with adjusted weighting parameters is also presented.

4.1 SLS Validation

Before presenting the detailed performance analysis of the trained DRL agents, the behavior of the SLS is demonstrated to establish its alignment with theoretical expectations. Figure 4.1 illustrates the relationship between DTX/DRX ON duration and mean perceived throughput across 3-cell and 7-cell configurations. As expected, mean throughput increases with longer ON durations for all cells, reflecting the greater number of transmission opportunities available within each cycle. Figure 4.2 presents the CDFs of Reference Signal Received Power (RSRP) with and without interference modelling. When interference is enabled, the starting points of the CDFs shift leftward, indicating that lower RSRP values are observed due to signal degradation due to interference from neighbouring cells. These observations confirm that the SLS correctly captures fundamental network dynamics, providing a reliable foundation for evaluating the DRL-based optimization framework.

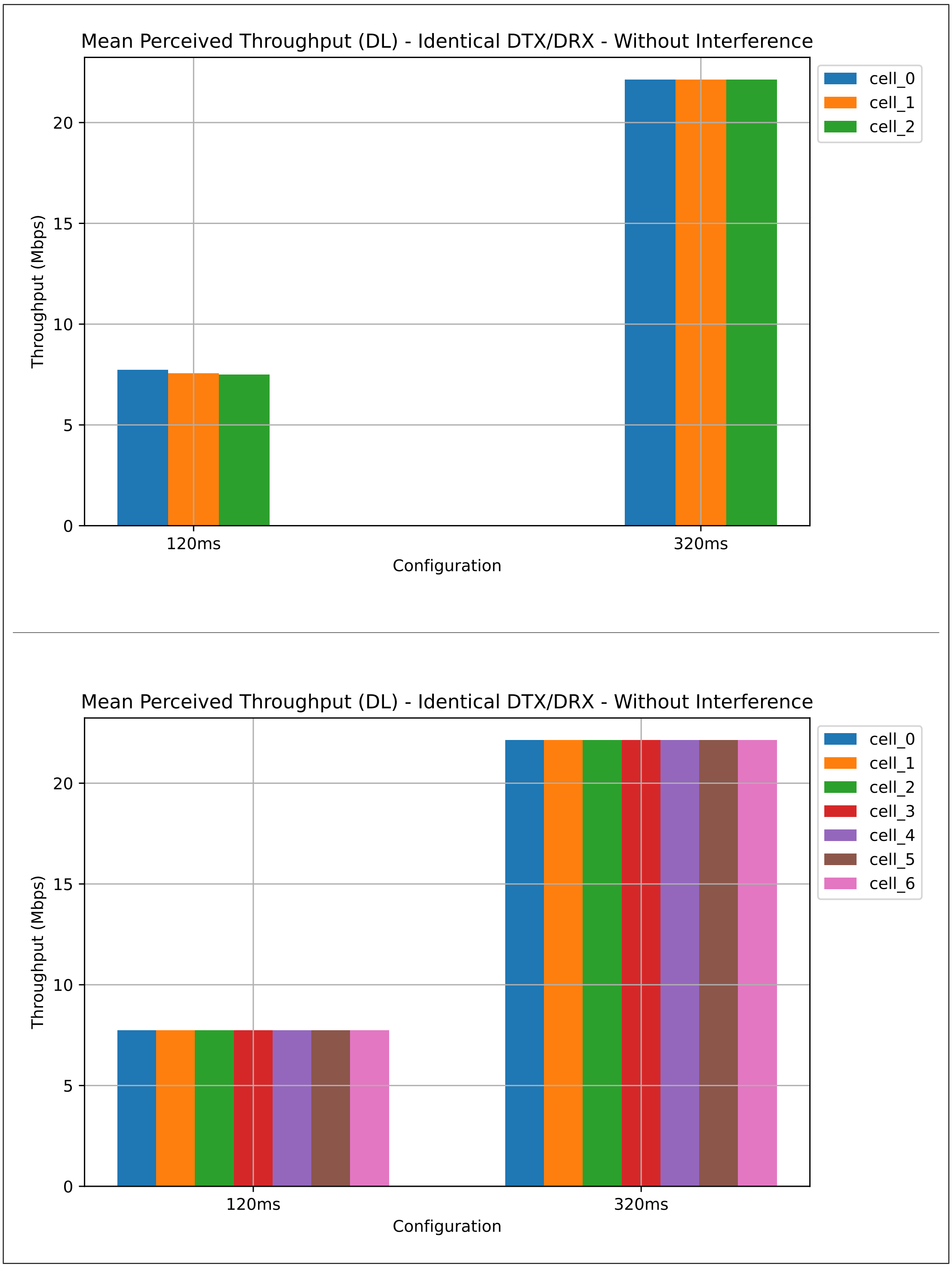

Figure 4.1 *Mean Perceived Throughput Comparison for Different DTX/DTX ON Durations.*

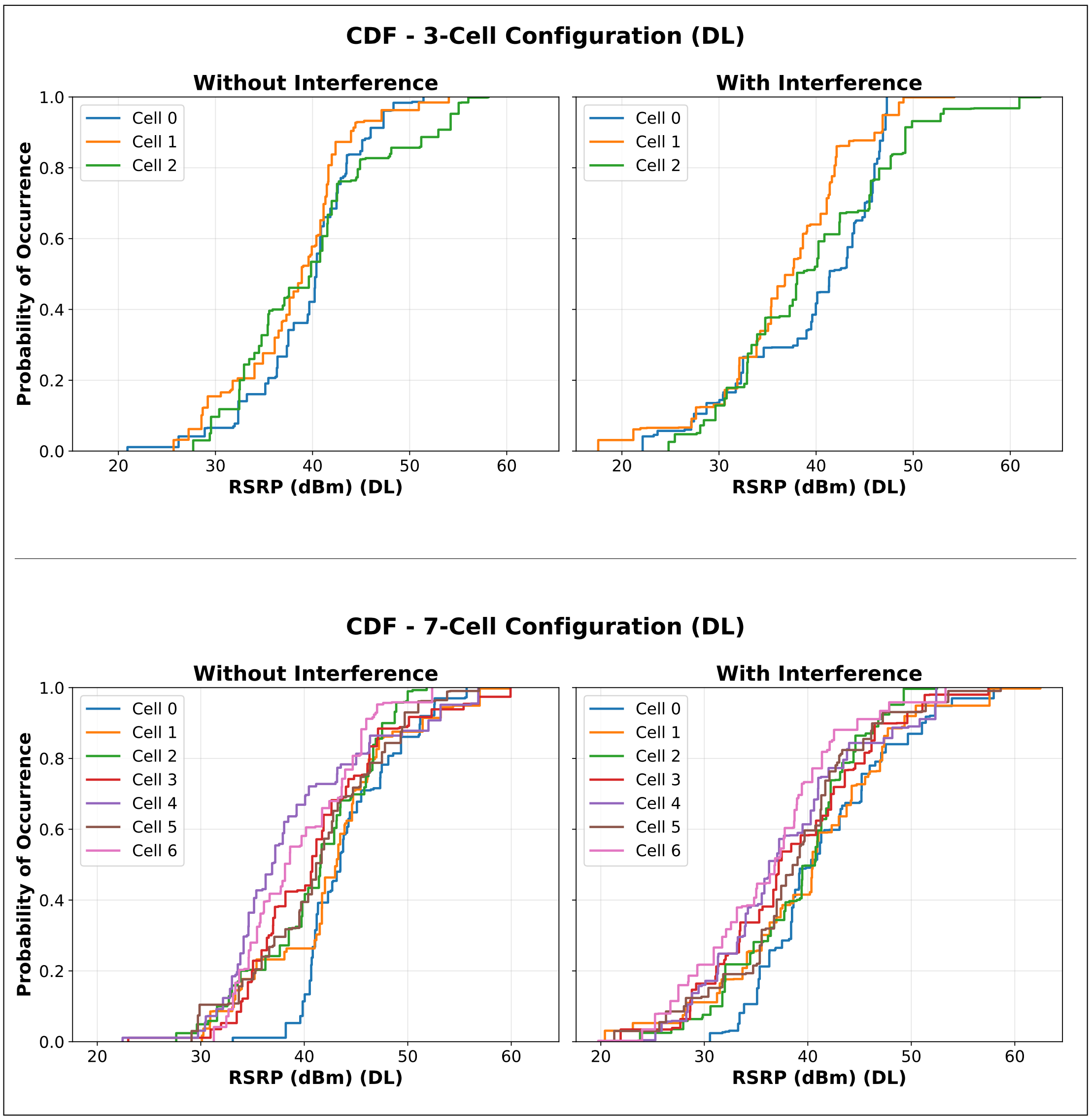

Figure 4.2 *RSRP CDFs with Interference Enabled and Disabled.*

4.2 Reward Function 1 Results

As described in the Section 3.3.3, Reward Function 1 employs a reward formulation based on EE ratio. The optional idle penalty was enabled for this experiment.

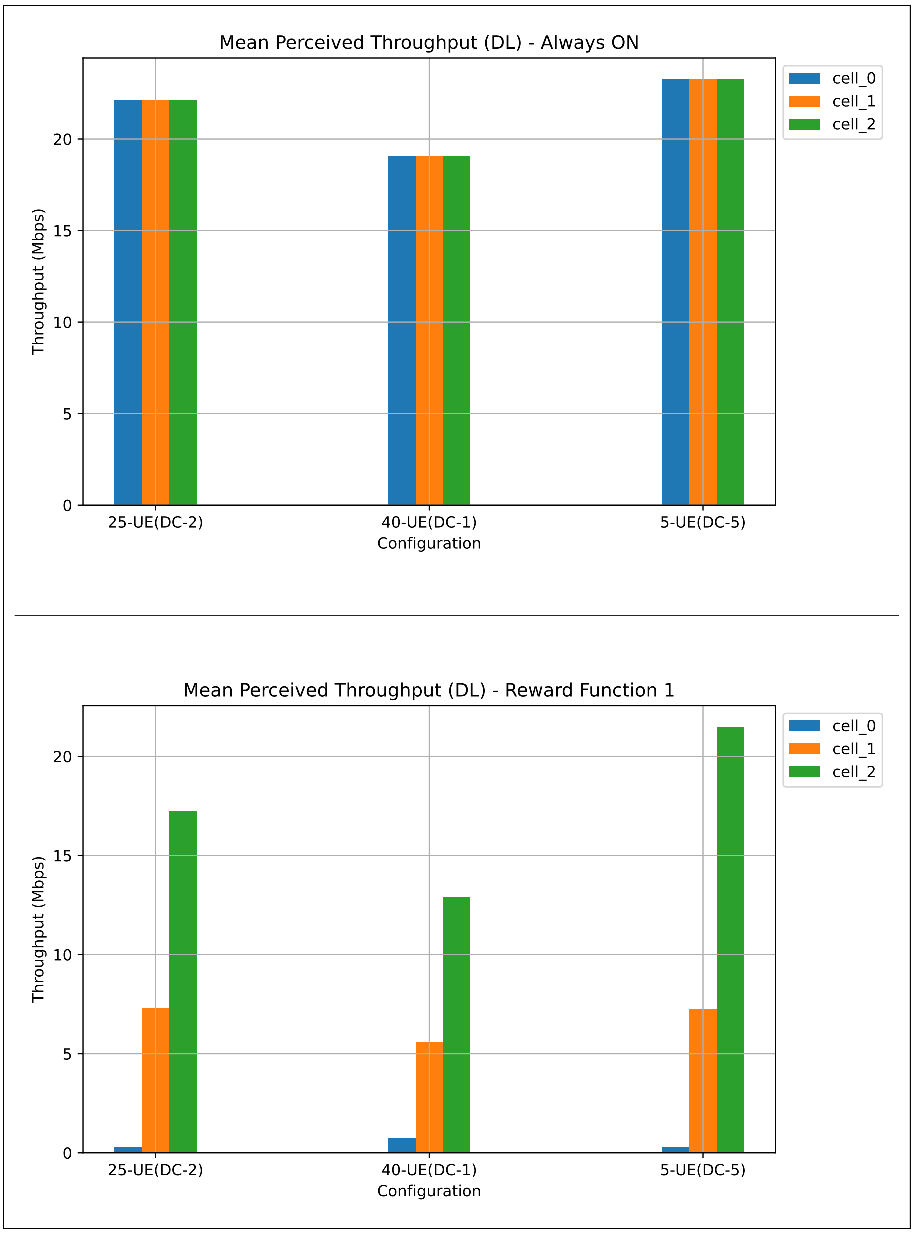

Figure 4.3 *Mean Perceived Throughput Comparison for Reward Function 1.*

Figure 4.3 presents the mean perceived throughput comparison between the Always-ON baseline and the trained agent across all deployment configurations. The top subplot shows the Always-ON baseline performance for each deployment

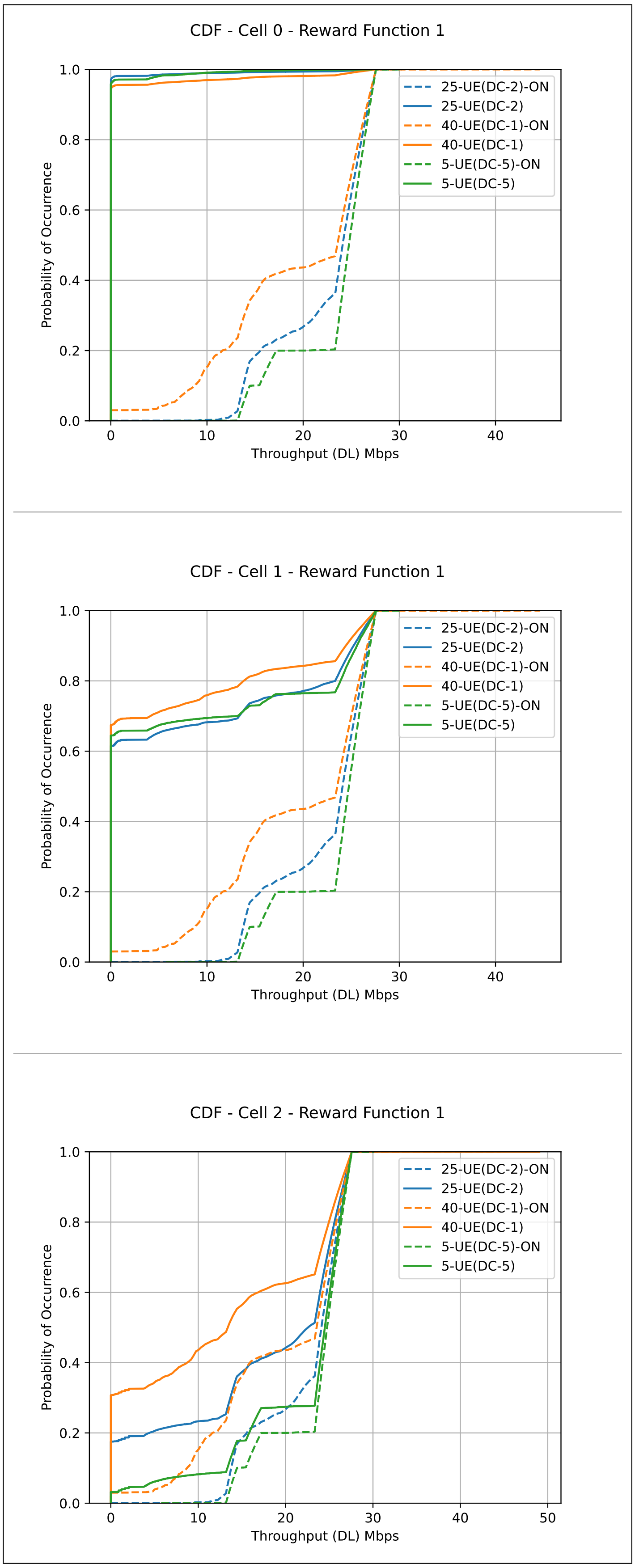

Figure 4.4 *Perceived Throughput CDFs for Reward Function 1.*

configuration, while the bottom subplot shows the agent's performance when inferenced on each deployment configuration. The throughput CDFs in Figure 4.4 provide detailed distribution information for each cell across all deployment configurations. Each of the three subplots corresponds to one cell (Cell_0, Cell_1, Cell_2), with separate curves for each deployment configuration. Dotted lines represent the Always-ON baseline while solid lines show the agent's performance, with consistent colour coding per deployment configuration.

As can be observed from the above plots, the mean perceived throughput performance has degraded significantly compared to the Always-ON scenario. The degradation is worst in Cell_0 as the mean perceived throughput observed is zero more than 90% of the time. The performance gets better in Cell_1 and Cell_2 as seen from the CDFs. However, the overall performance is not acceptable for real-world scenarios. The mean delay plots in Figure 4.8 and Figure 4.9, serves as confirmation for this sub-optimal performance

The sub-optimal performance observed in mean perceived throughput is explained by the mean static power consumption plots (Figure 4.5 and Figure 4.6) and the predicted actions of the agent (Figure 4.6). Cell_0 opts to remain OFF for the majority of time across all configurations, with the percentage of time spent in the ON state decreasing for sparser traffic scenarios. Cell_1 predominantly selects an ON duration of approximately 100ms (roughly 0.3125 times the cycle duration), again showing lower utilisation of higher ON durations in lighter traffic conditions. Cell_2 favours an ON duration of 300ms (approximately 0.9375 times the cycle duration) for most configurations, though counterintuitively, the percentage of time spent at lower ON durations decreases rather than increases for sparser traffic scenarios.

This aggressive energy saving behaviour stems from fundamental limitations in the reward function design identified during preliminary training experiments. The reward formulation only considers successfully transmitted packets in the numerator, providing no direct visibility into packet drops. Packets that exceeded their delay budget and were discarded had no explicit representation in the reward signal. Consequently, the agent learned policies that aggressively minimised active durations to reduce energy consumption, leading to excessive packet drops due to insufficient transmission opportunities. Without explicit feedback about these dropped packets, the agent lacked the information necessary to understand the QoS implications of its actions. The agent essentially learned to optimize the reward function itself rather than achieving the intended network performance objectives, demonstrating a clas-

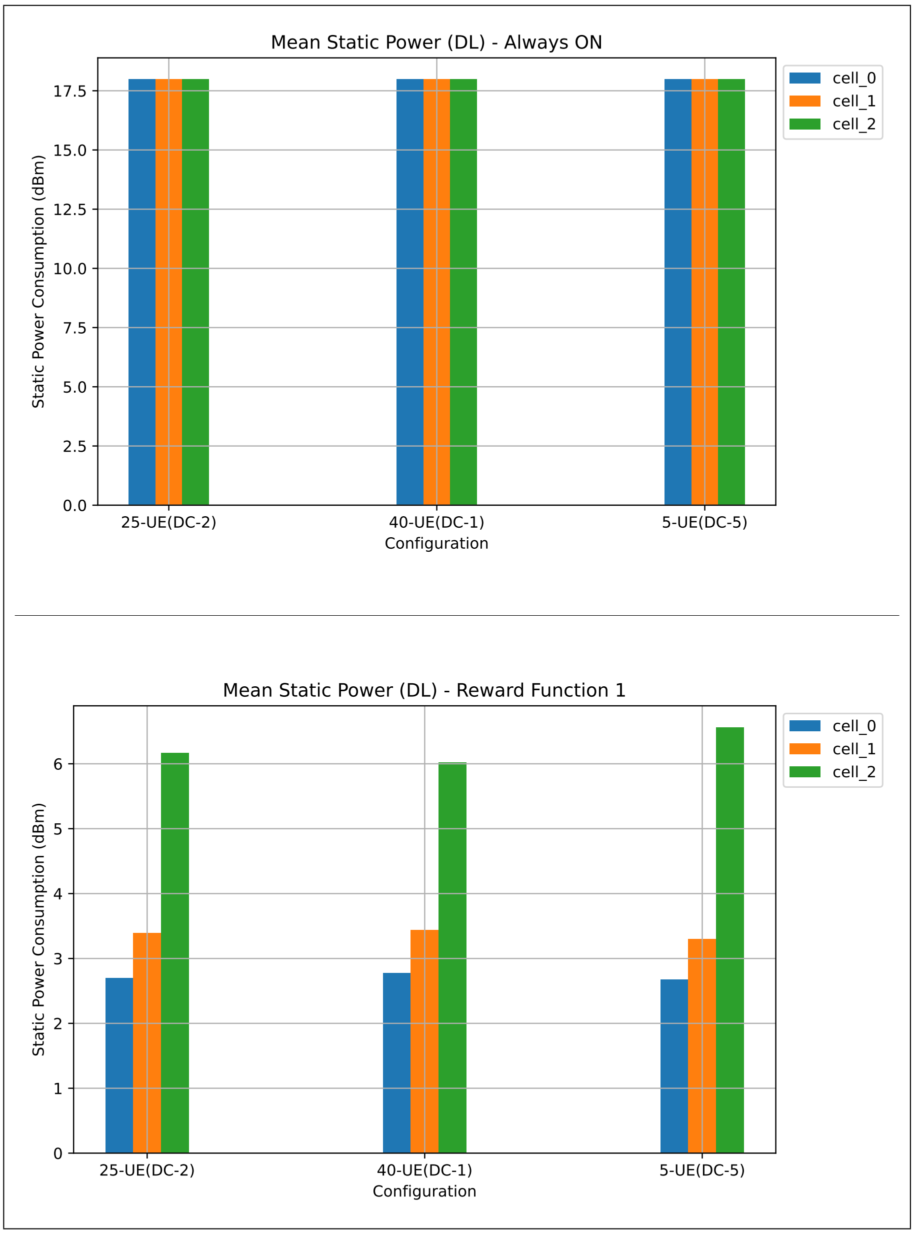

Figure 4.5 *Mean Static Power Consumption Comparison for Reward Function 1.*

sic reward misalignment problem where high reward values were obtained despite poor actual network performance.

4.3 Reward Function 2 Results

As described in the Section 3.3.3, the second reward function incorporated modifications to address the reward misalignment observed in Reward Function 1. The objective of this formulation is to force the agent to minimize dropped packets and thus better balance NES and QoS performance.

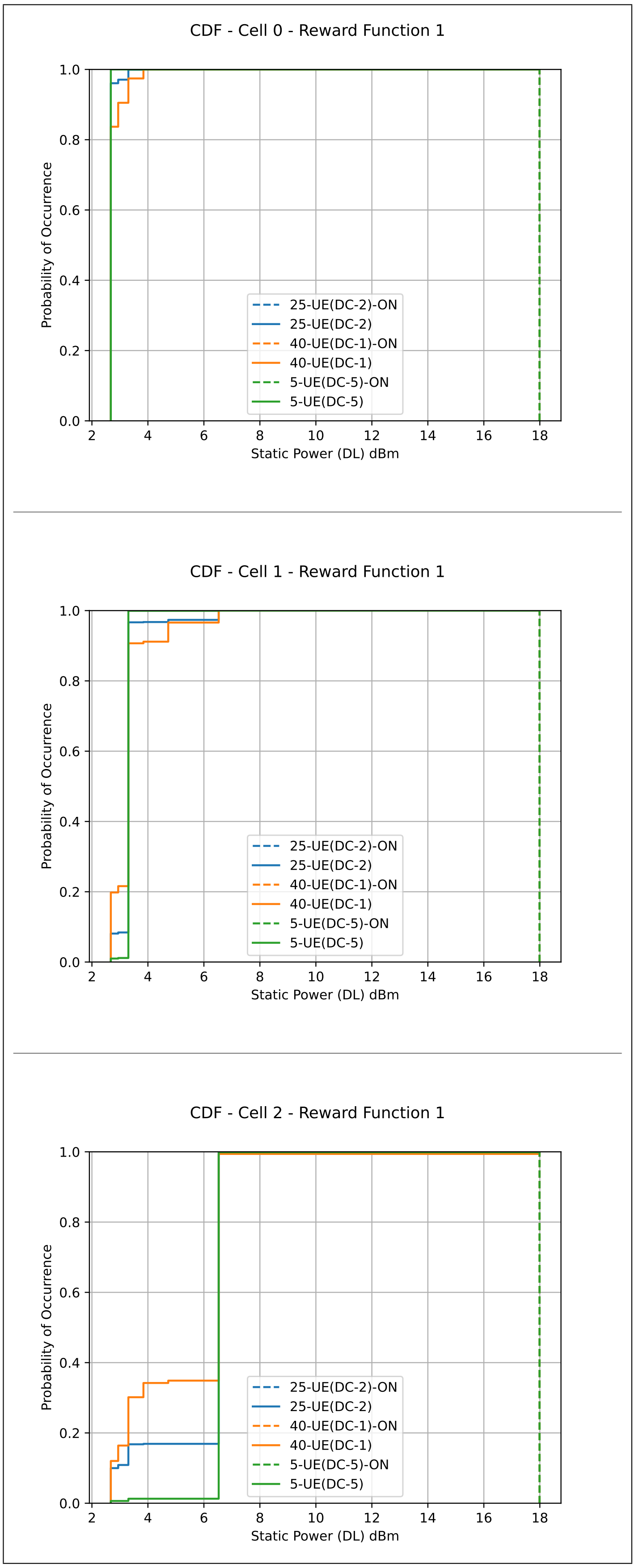

Figure 4.6 *Static Power Consumption CDFs for Reward Function 1.*

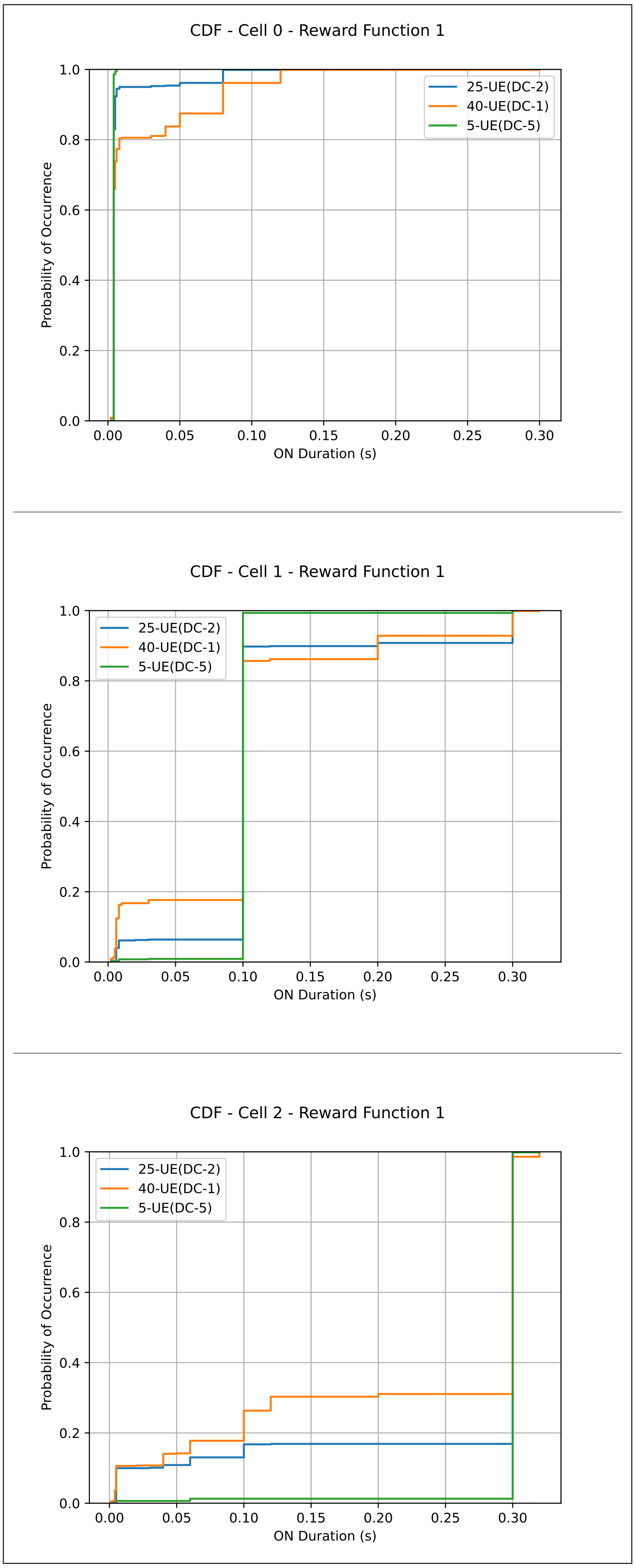

Figure 4.7 *Predicted Actions CDFs for Reward Function 1.*

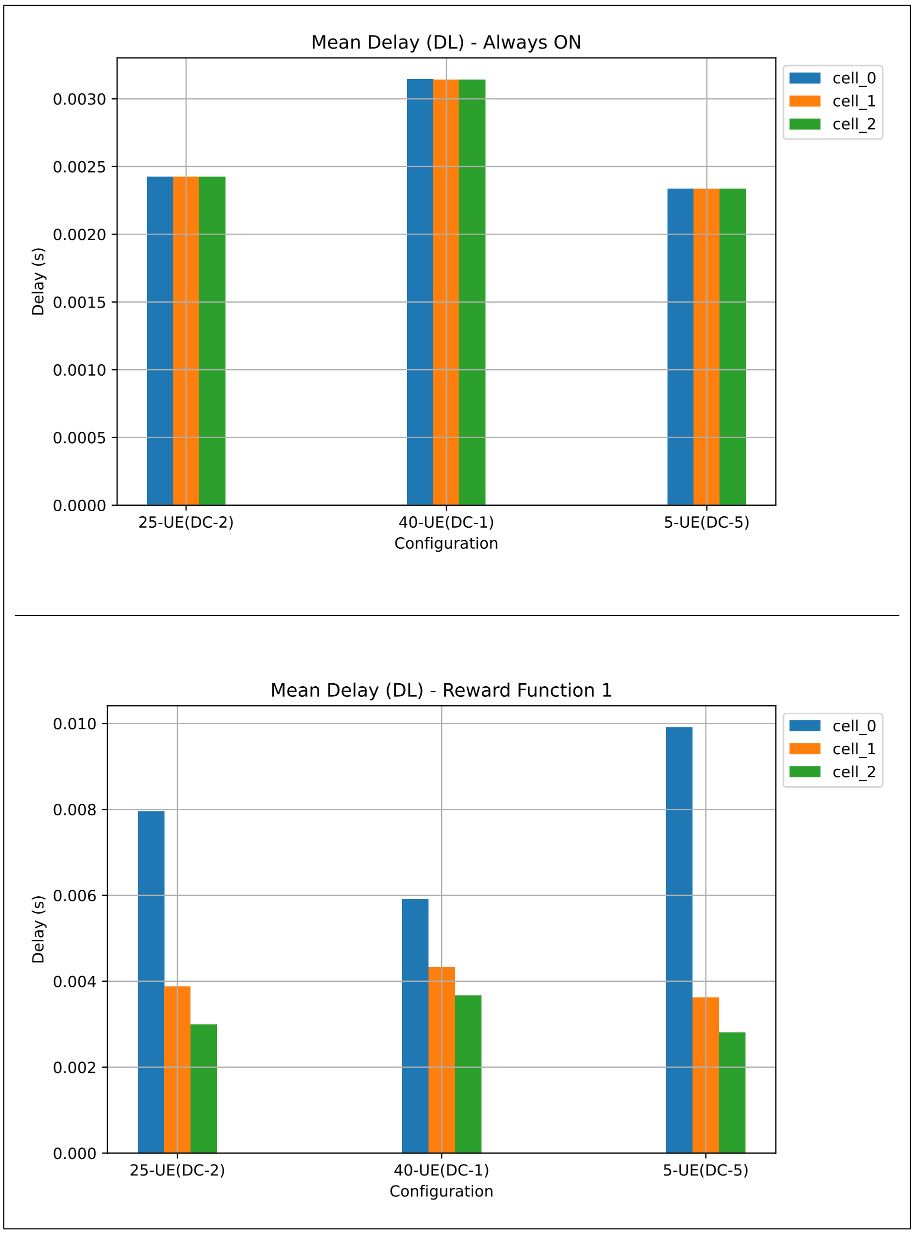

Figure 4.8 *Mean Delay Comparison for Reward Function 1.*

As can be observed from the plots in Figure 4.10 and Figure 4.11, the mean perceived throughput performance has degraded significantly compared to the Always-ON scenario. The performance here in Cell_0 is slightly better when compared to the results of Reward Function 1. However, the performance has become worse for Cell_1 and Cell_2 compared to the Reward Function 1. The overall performance is counter-intuitive to the expectation signalling deeper issues in reward formulation in addition to the packet drops. Again, the mean delay plots in Figure 4.15 and Figure 4.16, backs these observations.

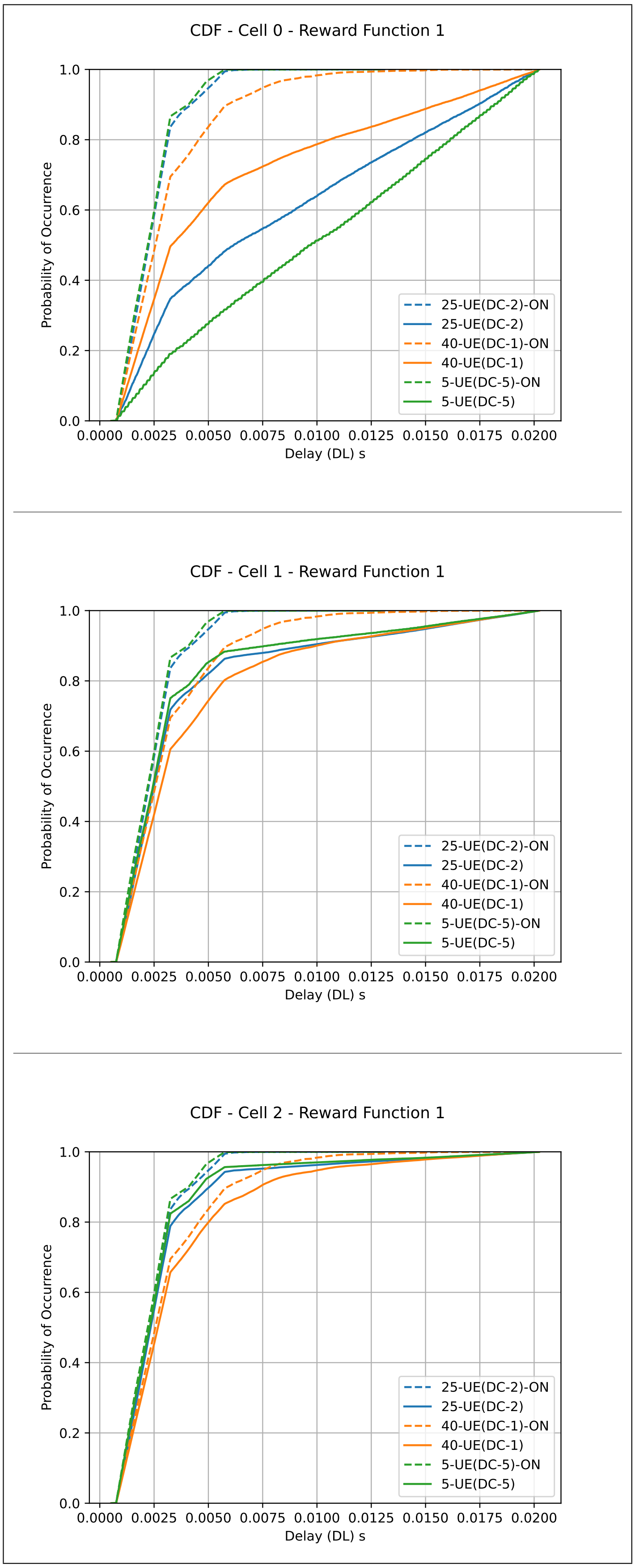

Figure 4.9 *Delay CDFs for Reward Function 1.*

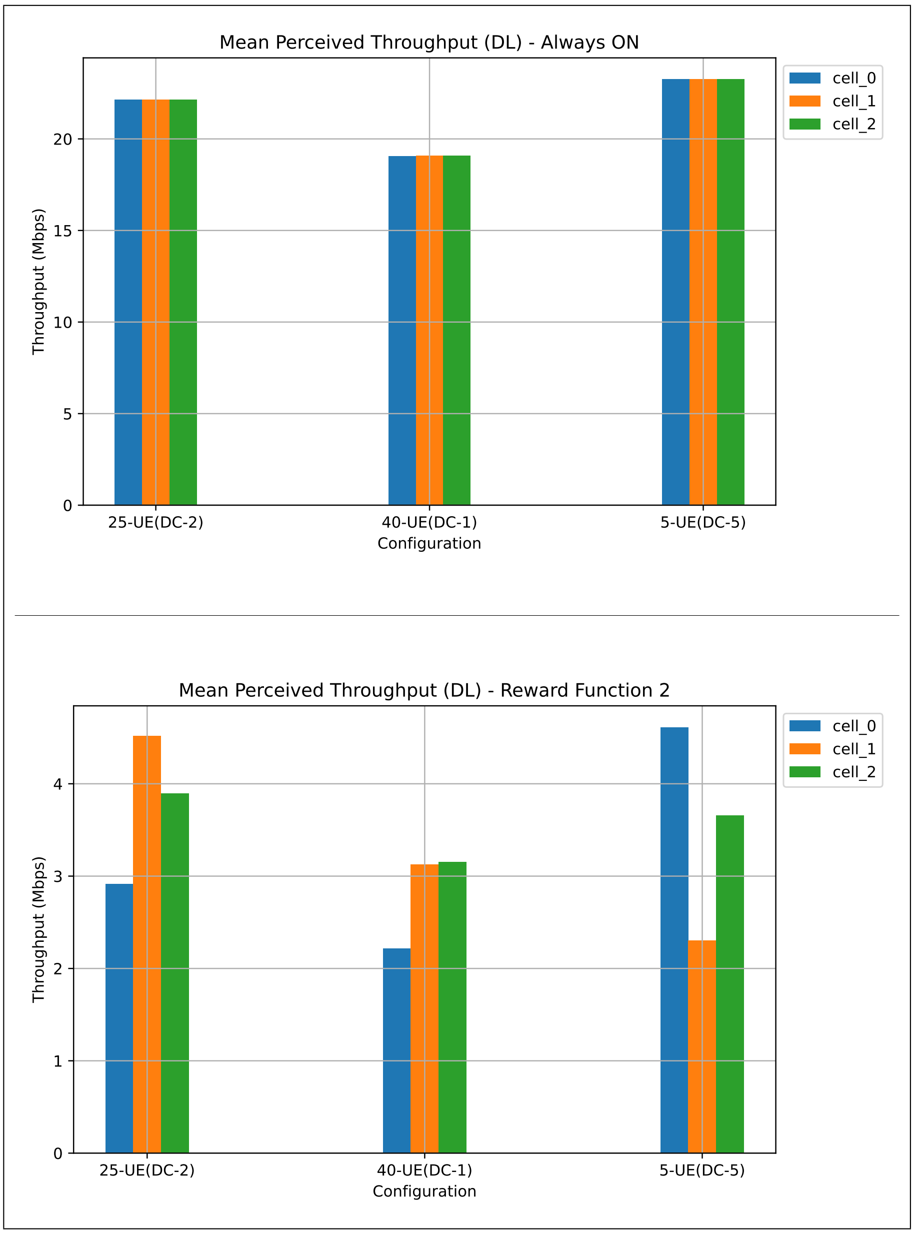

Figure 4.10 *Mean Perceived Throughput Comparison for Reward Function 2.*

The trained agent continues to exhibit overly aggressive energy saving behaviour, though with different patterns compared to Reward Function 1 as confirmed by the mean static power consumption plots (Figure 4.12 and Figure 4.13) and the predicted actions of the agent (Figure 4.14). Analysis of the predicted ON duration distributions reveals that Cell_0 opts to either remain OFF or stay at ON durations below 200ms for the majority of time across all configurations. The percentage of time spent in OFF state or using very low ON durations (roughly below 30ms) increases for sparser traffic scenarios, representing intuitive adaptation. However, the pattern reverses when comparing the percentage of time spent below 200ms, suggesting the agent fails to appropriately adapt to reduced traffic loads. Cell_1

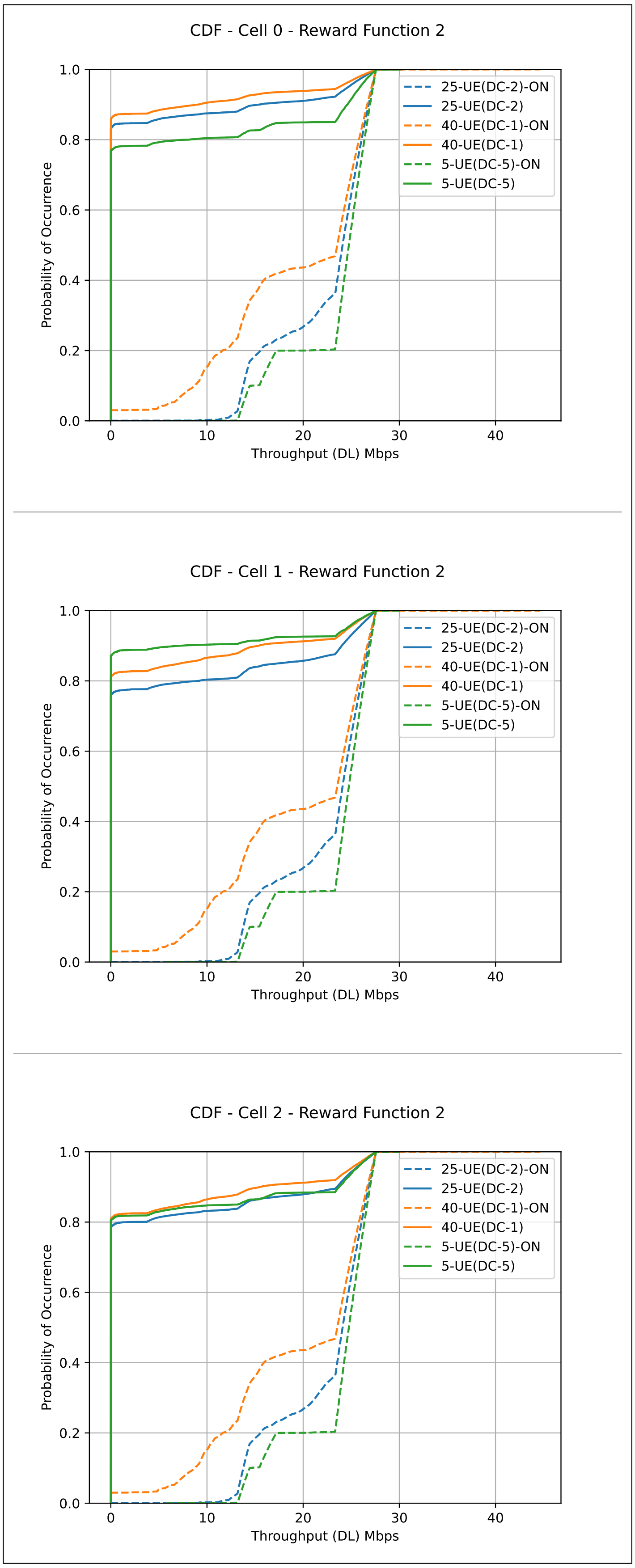

Figure 4.11 *Perceived Throughput CDFs for Reward Function 2.*

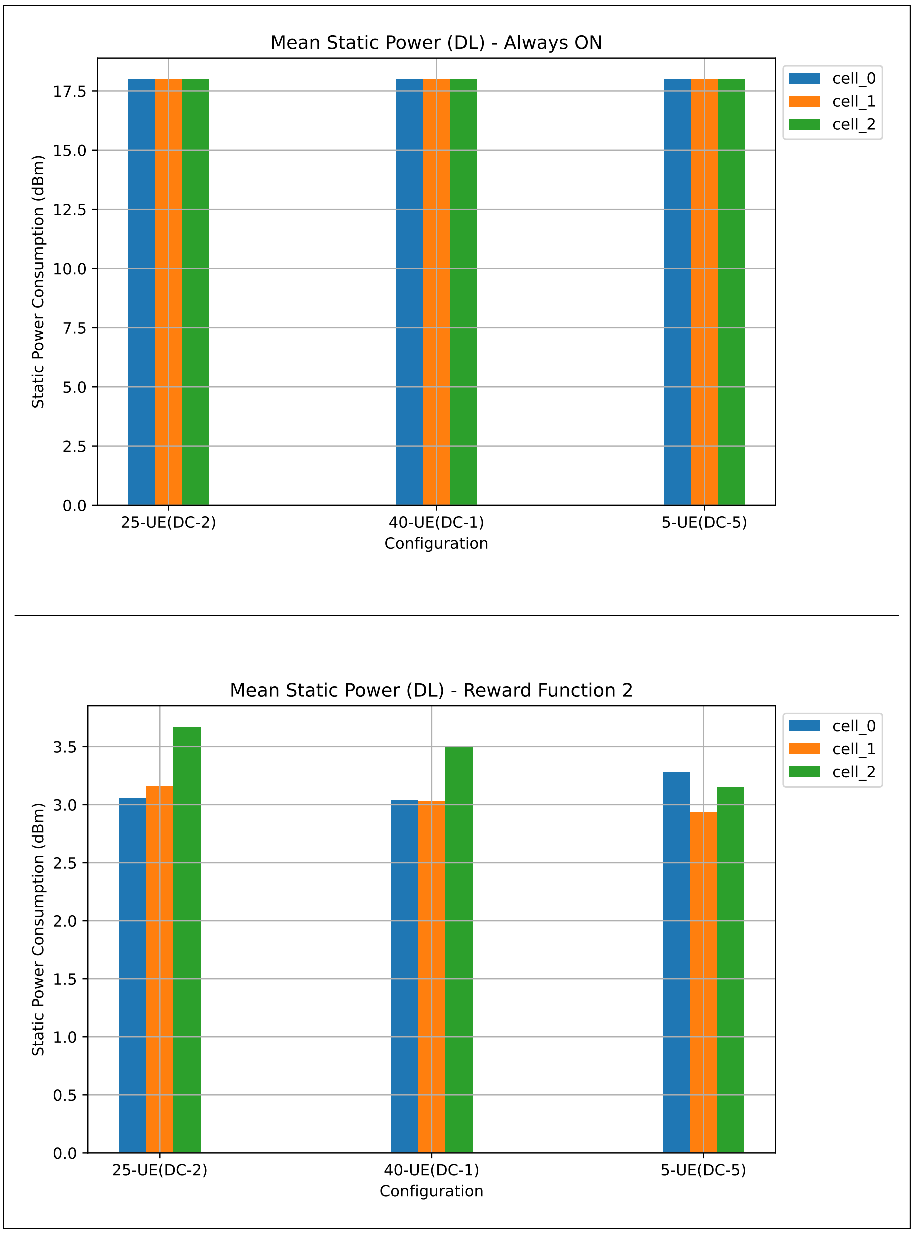

Figure 4.12 *Mean Static Power Consumption Comparison for Reward Function 2.*

demonstrates similar behaviour to Cell_1, selecting either OFF or low ON durations (below 200ms) predominantly, with no clear pattern in the percentage of time spent in these states based on deployment configuration. Cell 2 behaves intuitively more often compared to Cell_0, as the percentage of time spent in OFF or low ON duration states increases for sparser traffic scenarios. The overall behaviour of the agent is too aggressive on energy saving which leads to degraded QoS performance.

While Reward Function 2 incorporated dropped packets into the reward signal to address the QoS visibility issues in Reward Function 1, the results demonstrate that fundamental limitations persisted in the ratio-based reward structure. The critical

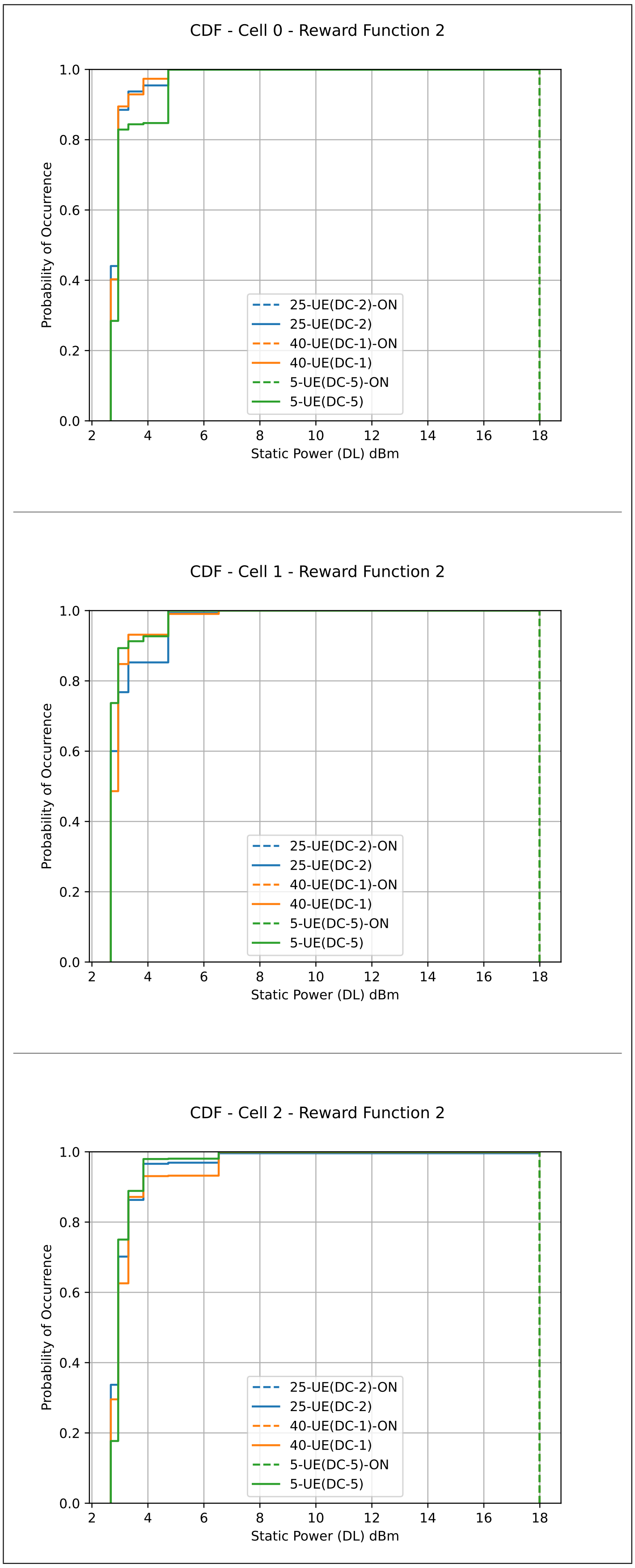

Figure 4.13 *Static Power Consumption CDFs for Reward Function 2.*

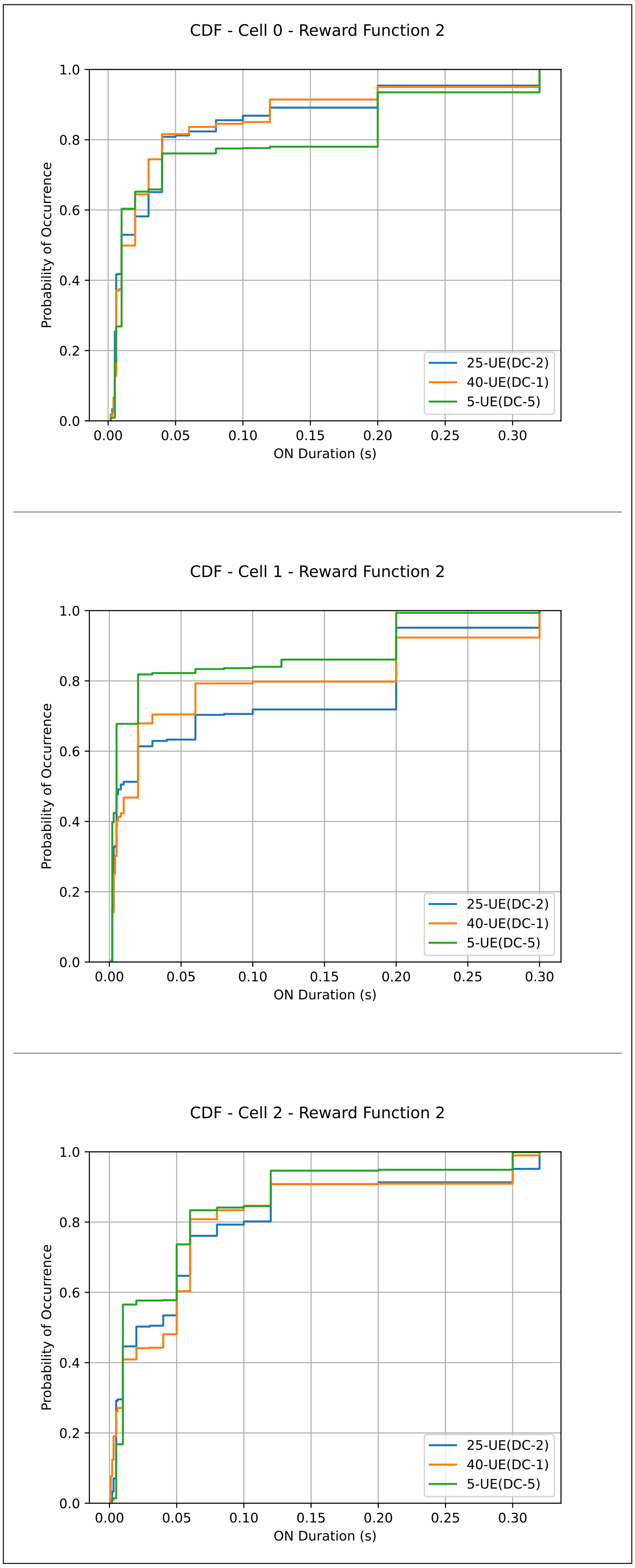

Figure 4.14 *Predicted Actions CDFs for Reward Function 2.*

flaw surfaces during periods when no traffic is present in the network. In such scenarios, both total packets and dropped packets equal zero, causing the reward numerator to become zero regardless of the denominator value. Consequently, the reward evaluates to zero whether the agent keeps cells active (wasting energy) or puts them to sleep (conserving energy optimally). This creates a null gradient situation where the agent receives identical reward signals for dramatically different behaviours. Thus, the reward function fails to provide meaningful differentiation between sleep and wasteful active states during idle periods. This ambiguity led to inconsistent and sub-optimal policy learning, with the agent unable to properly balance NES with QoS requirements across varying traffic conditions.

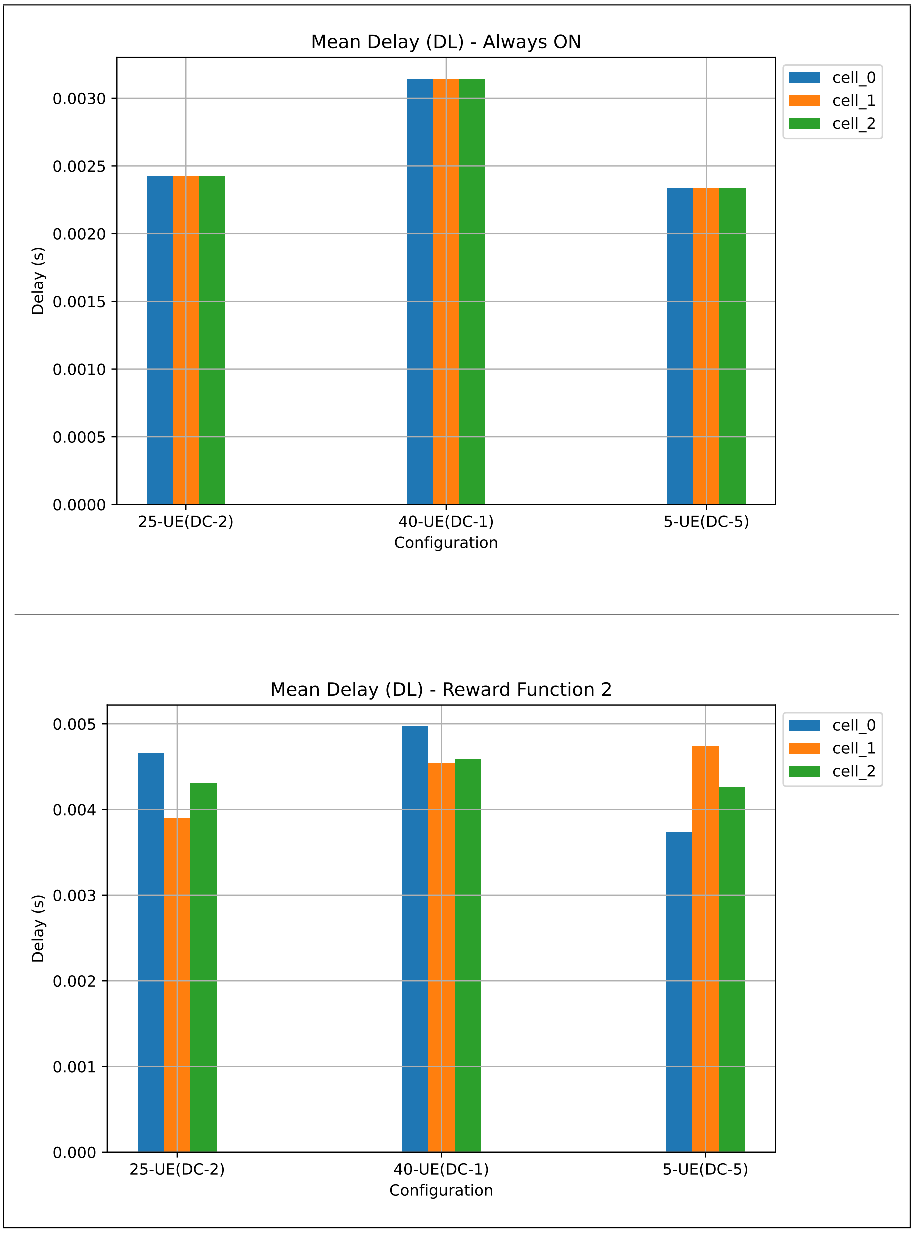

Figure 4.15 *Mean Delay Comparison for Reward Function 2.*

4.4 Reward Function 3 Results

As described in Section 3.3.3, the Reward Function 3 incorporated lessons learned from the previous iterations. This formulation uses a conditional linear reward structure with separate reward components for high-backlog and low-backlog states, better aligning the optimization objectives with actual network performance requirements.

As can be observed from the above plots (Figure 4.17 and Figure 4.18), the mean perceived throughput performance again shows a degradation compared to the Always-ON scenario. However, the overall performance is better when compared to that of both Reward Functions 1 and 2. The mean perceived throughput observed is zero more than 60% of the time in Cell_0. The performance gets better in Cell_1 and Cell_2 as seen from the CDFs. Again, the mean delay plots in Figure 4.22 and Figure 4.23, mirror these observations.

Analysis of the predicted ON duration distributions (Figure 4.21) and the mean static power consumption plots (Figure 4.19 and Figure 4.20) reveals distinct adaptive patterns across the three cells. Cell_0 predominantly selects an ON durations of 120ms (approximately 0.375 times the cycle duration) for most configurations. The percentage of time spent at ON durations below 120ms increases for sparser traffic scenarios, indicating intuitive adaptation. Cell 1 exhibits a bimodal distribution, mostly opting to stay ON for either 80ms (0.25 times the cycle duration) or 320ms (cycle duration) across all configurations. However, counter-intuitively, the percentage of time spent at higher ON durations increases for sparser traffic scenarios, suggesting the agent prioritises maintaining capacity even under reduced load. Cell 2 demonstrates another bimodal pattern, mostly selecting either 100ms (0.3125 times the cycle duration) or 300ms (0.9375 times the cycle duration) for all configurations. The percentage of time spent at the 100ms ON duration increases for sparser traffic scenarios, representing intuitive energy saving adaptation. The overall behaviour of the agent the agent has becomes less aggressive in energy saving and starting to prioritize QoS requirements as well.

These results demonstrate that Reward Function 3 achieves substantially better performance compared to both Reward Function 1 and Reward Function 2, signalling that the reward formulation is converging towards an effective balance between EE and QoS objectives. The state-dependent reward structure successfully

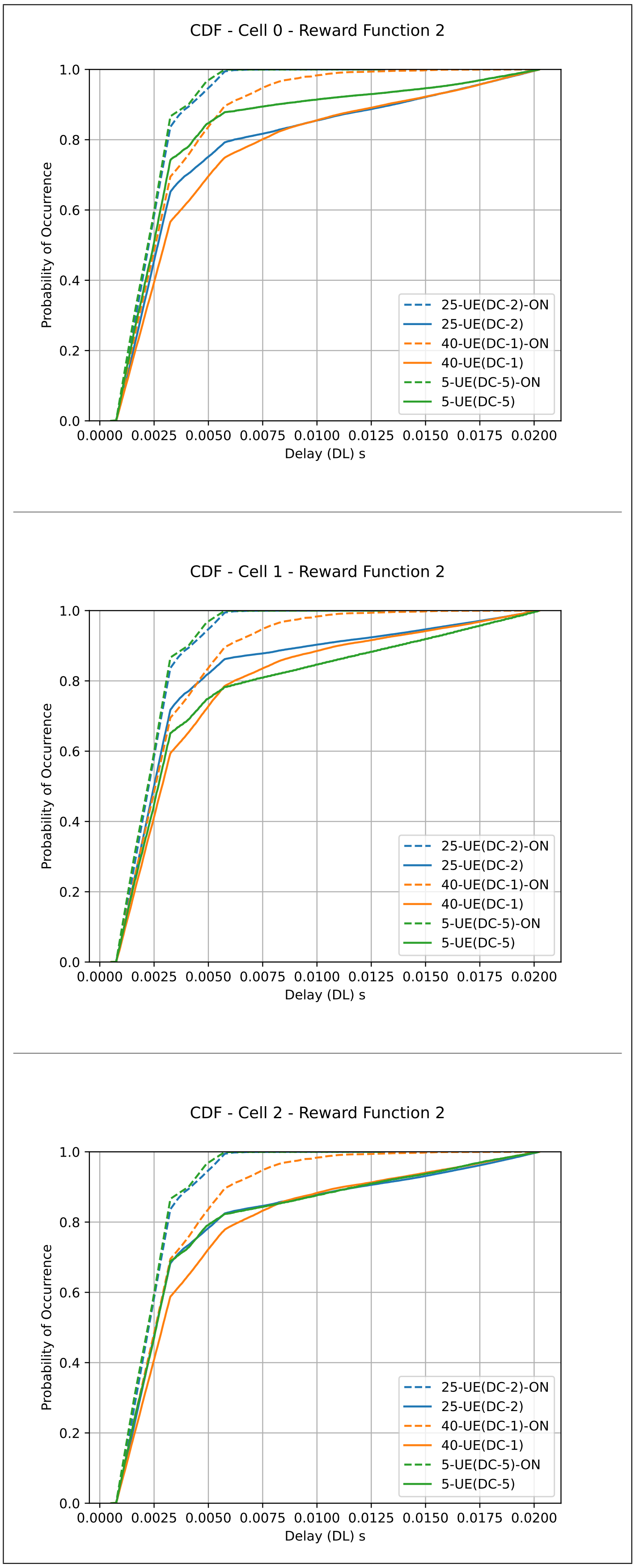

Figure 4.16 *Delay CDFs for Reward Function 2.*

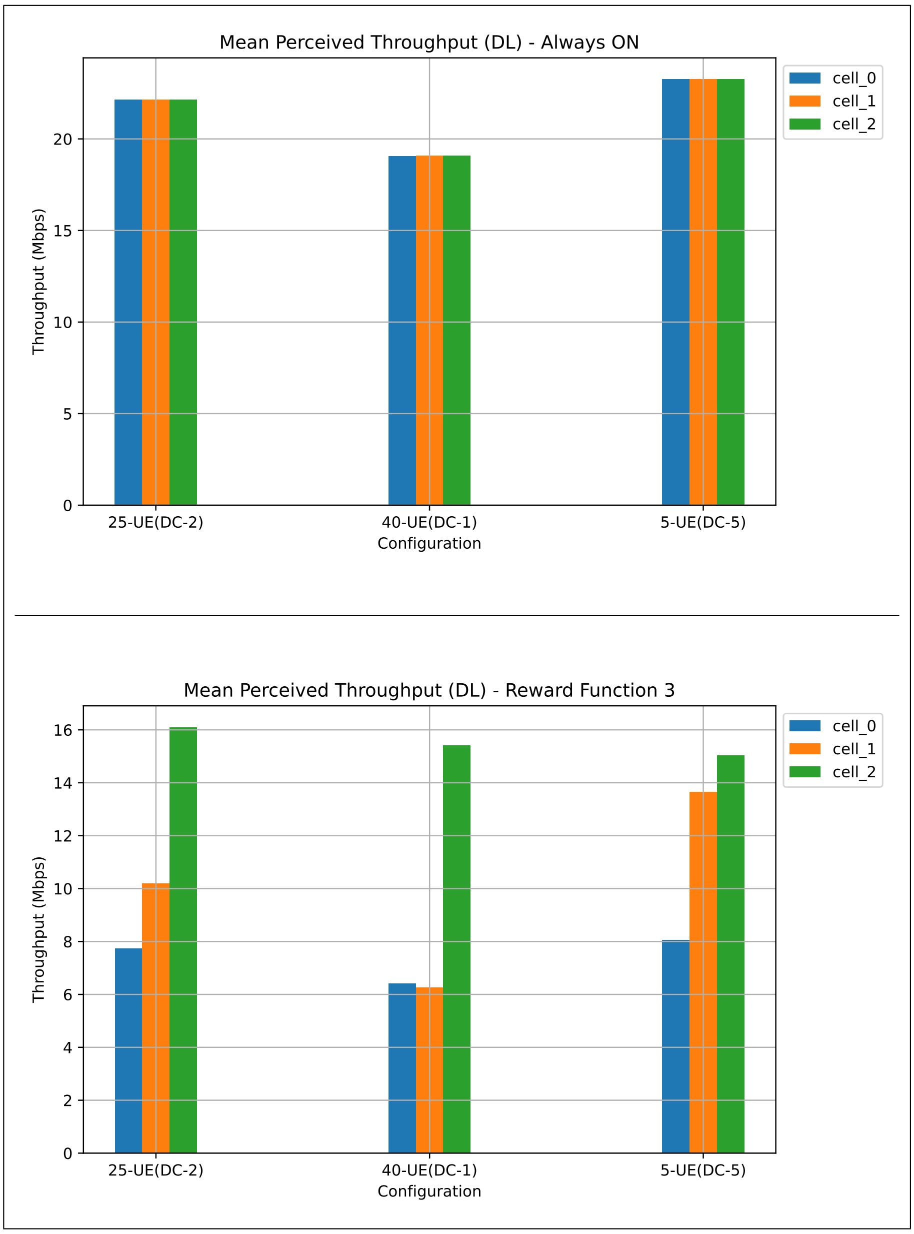

Figure 4.17 *Mean Perceived Throughput Comparison for Reward Function 3.*

addresses the fundamental limitations observed in the previous formulations, enabling the agent to learn more nuanced policies that adapt appropriately to varying network conditions. However, the counter-intuitive behaviour observed in Cell 1 and the remaining aggressive tendencies suggest that further refinement of the reward function parameters may yield additional improvements. This motivates exploration of modified weighting parameters to fine-tune the trade-off between NES and QoS requirements.

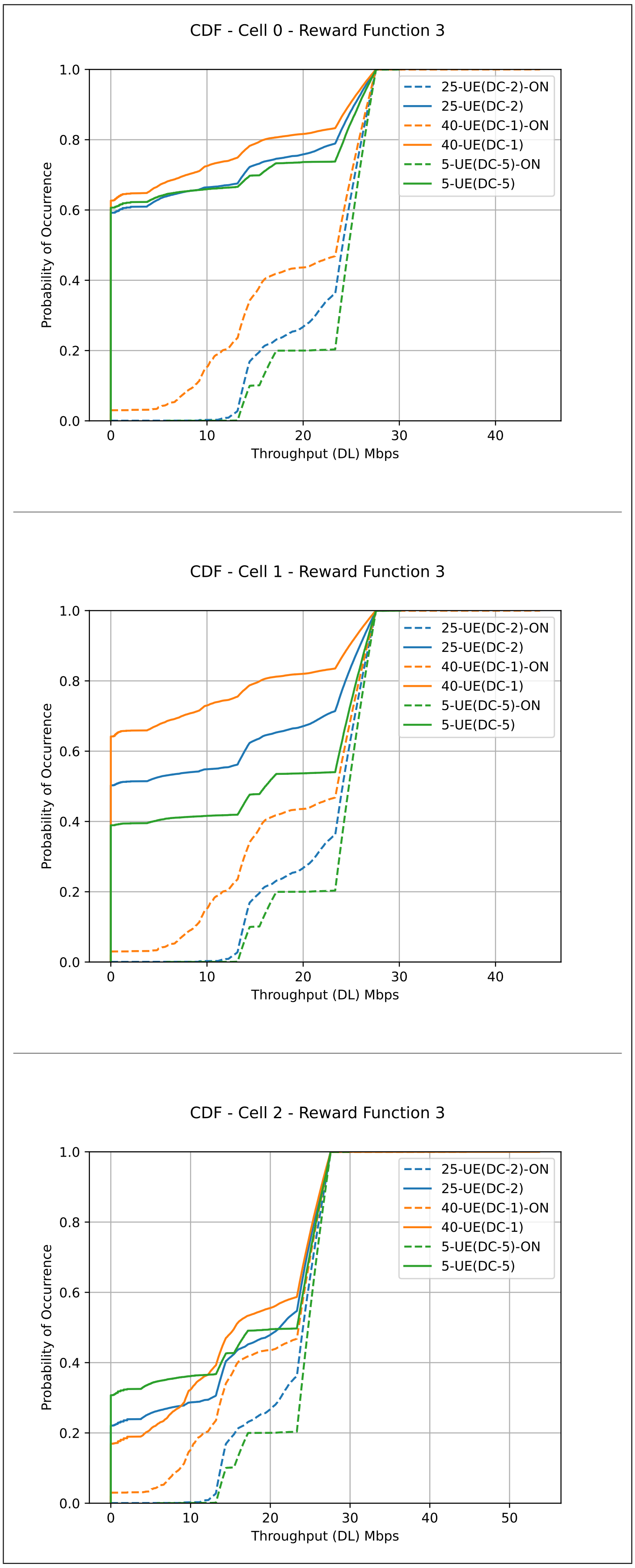

Figure 4.18 *Perceived Throughput CDFs for Reward Function 3.*

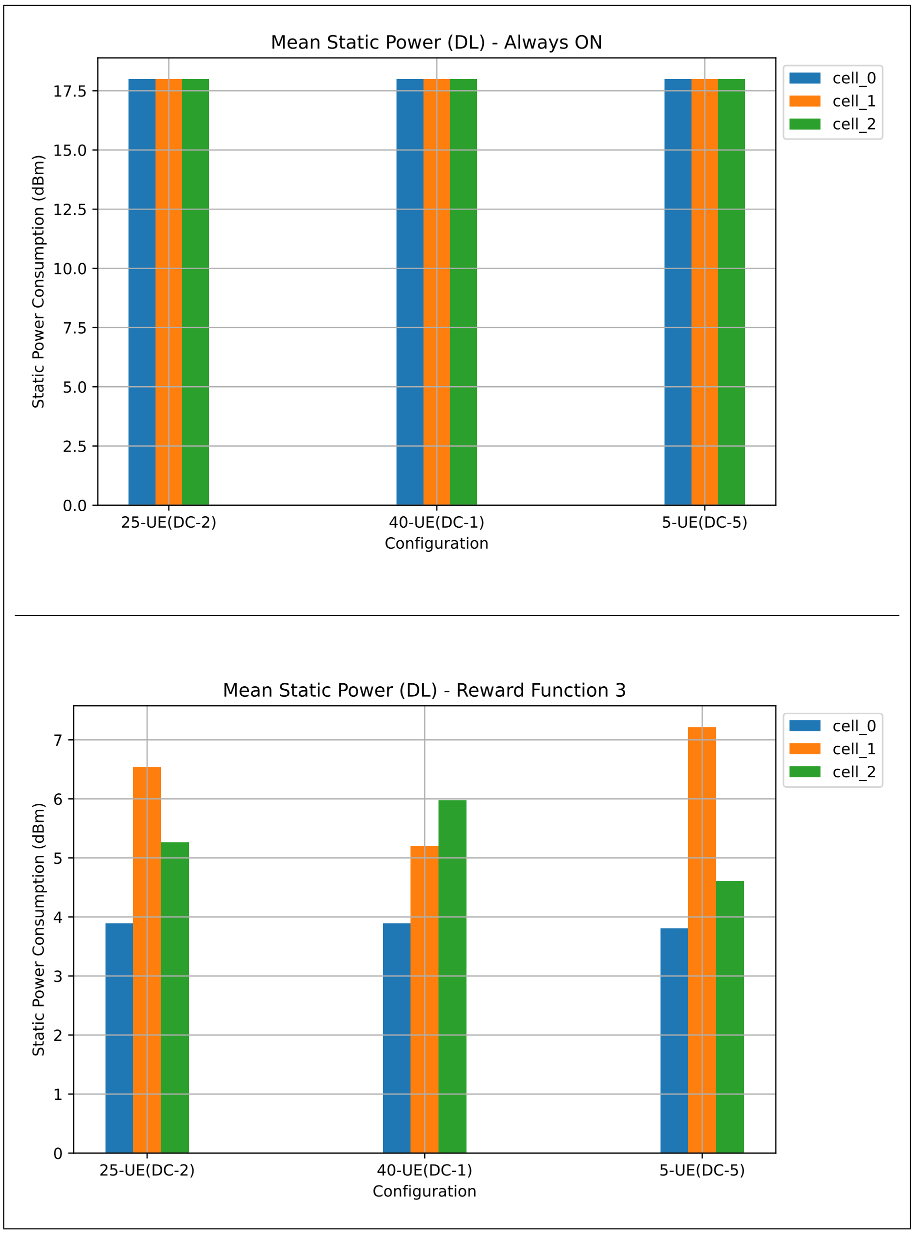

Figure 4.19 *Mean Static Power Consumption Comparison for Reward Function 3.*

4.5 Reward Function 3 with Modified Weights

To further explore the impact of reward function parameterization, a modified version of Reward Function 3 was evaluated with adjusted weighting parameters for the throughput and power components. The modified weights alter the relative importance of energy savings and QoS performance in the agent's decision-making process.

A slight modification is made when backlog is present here to the eq. (3.10). The

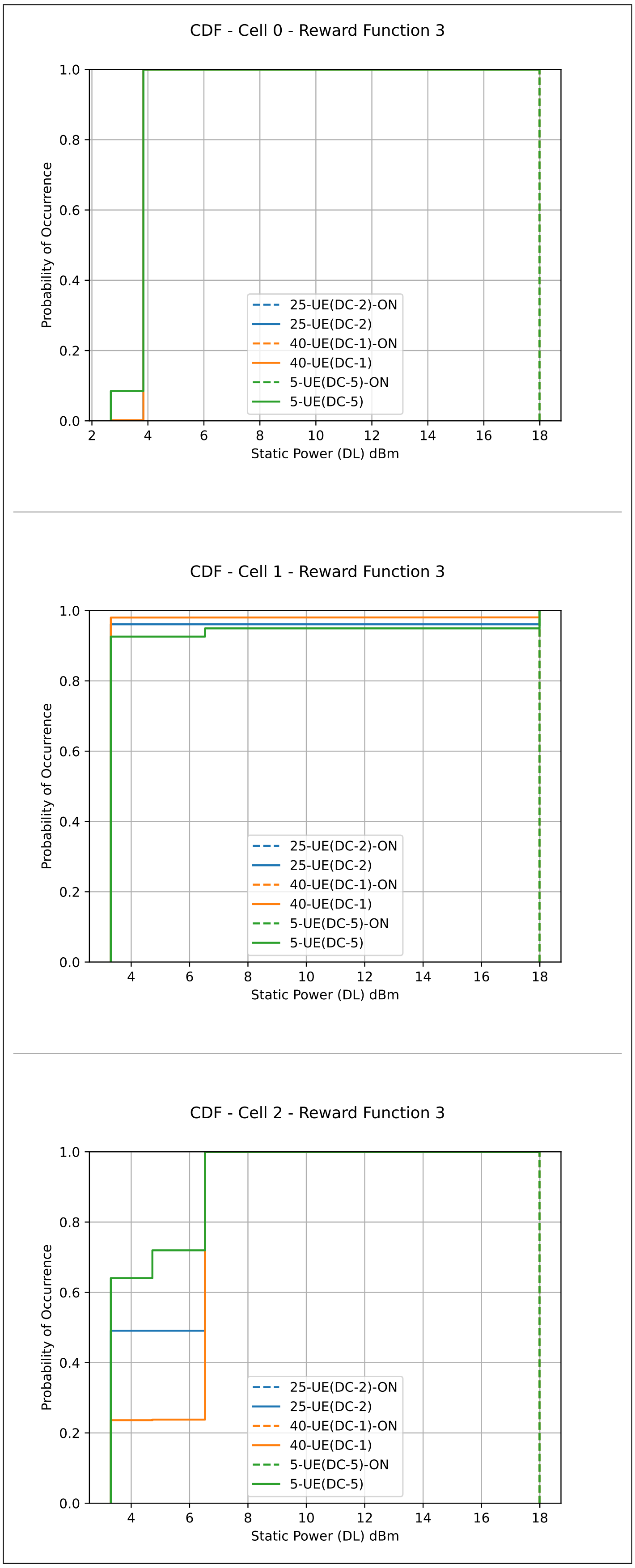

Figure 4.20 *Static Power Consumption CDFs for Reward Function 3.*

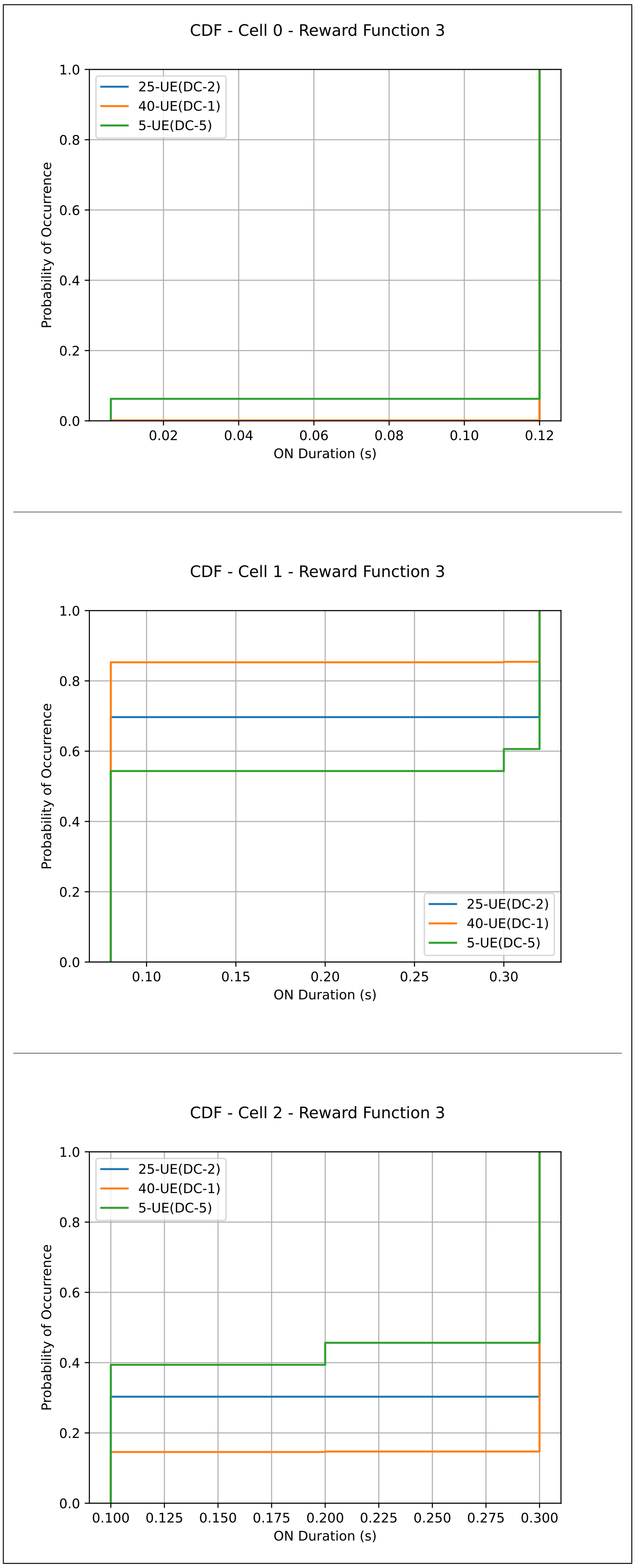

Figure 4.21 *Predicted Actions CDFs for Reward Function 3.*

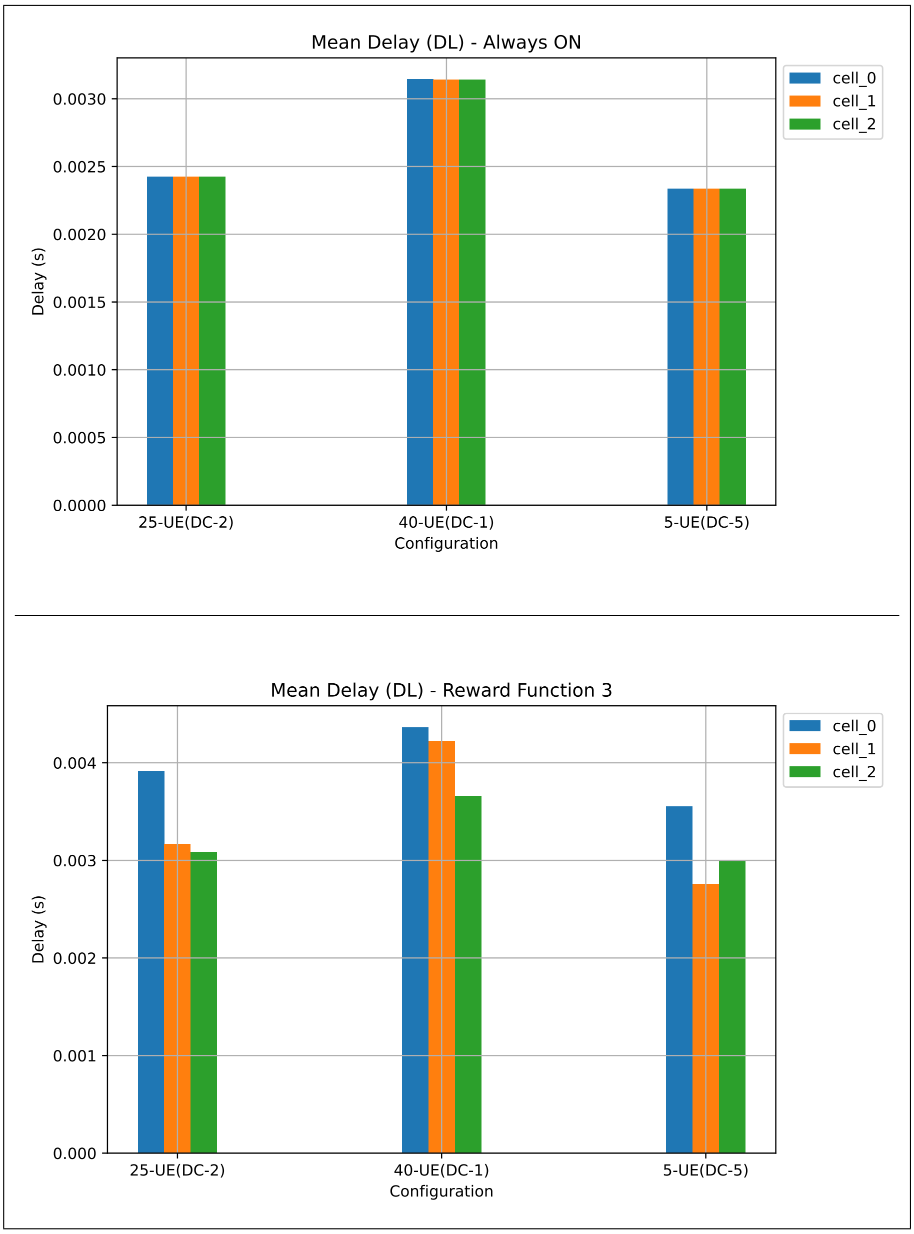

Figure 4.22 *Mean Delay Comparison for Reward Function 3.*

modified equation is:

$$R = w_{\text{service}} \cdot \frac{T_{\text{tot}}}{T_{\text{max}}} - w_{\text{energy}} \cdot \frac{E_{\text{tot}}}{E_{\text{max}}} - w_{\text{idle\_buf}} \cdot \text{Idle\_sum}, \tag{4.1}$$

where $\boldsymbol{w_{idle\_buf}}$ denotes a separate weight idle fraction when the buffer length is greater than zero. Earlier a similar weight was used for penalizing idling regardless of backlog status. This adjustment aims to reduce the aggressive energy-saving tendency observed in the original formulation. $\boldsymbol{w_{service}}$ was increased from 1.0 to 1.2, placing greater emphasis on successfully transmitted packets when backlog exists. This modification is aimed at forcing the agent to prioritize packet delivery.

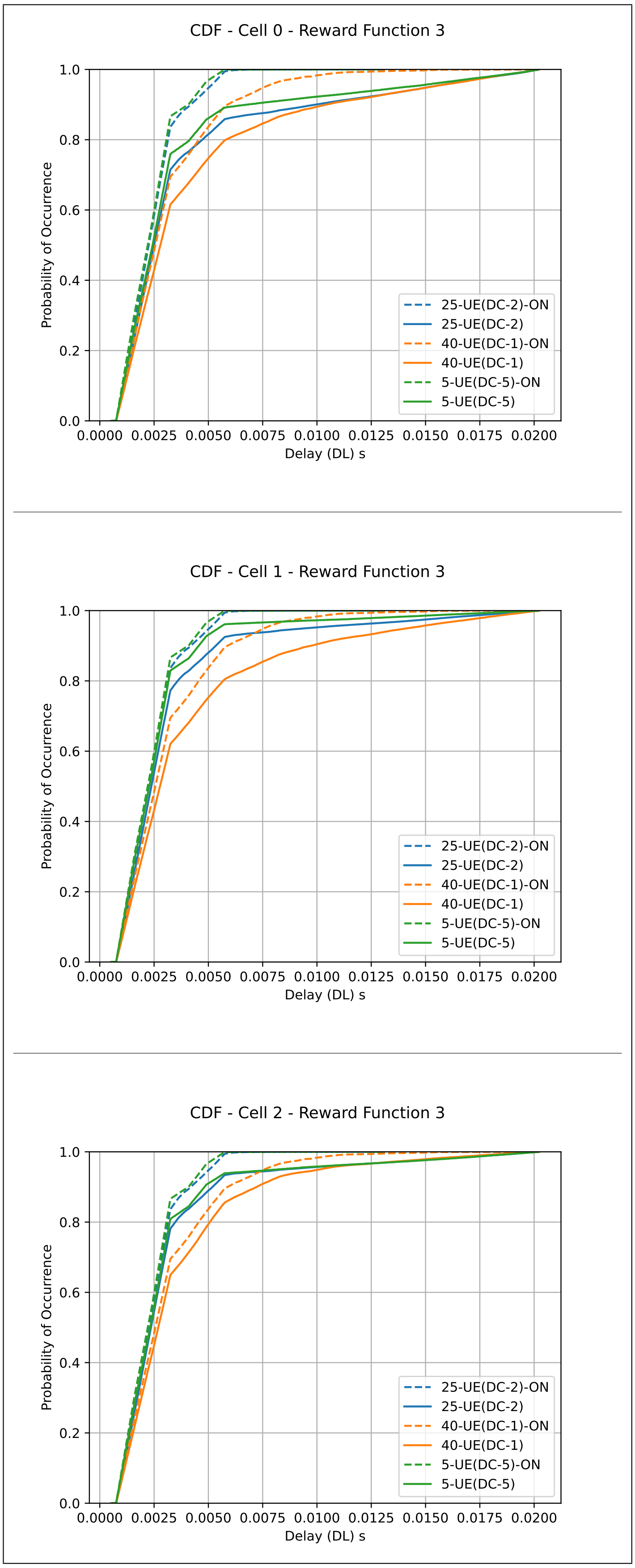

Figure 4.23 *Delay CDFs for Reward Function 3.*

Also, $\boldsymbol{w_{energy}}$ was reduced from 0.5 to 0.4, allowing the agent to consume more power. This modification is aimed at forcing the agent to refrain from aggressive power savings in order to achieve a higher reward. Table 4.1 summarizes the reward weight configuration used in this experiment. These parameter adjustments collectively aim to shift the agent's policy towards better QoS performance while preserving EE gains.

Parameter	**Value**
w_{sleep}	1.0
w_{idle}	0.5
$w_{\text{idle\_buf}}$	0.25
w_{service}	1.2
w_{energy}	0.4

Table 4.1 *Modified Reward Weight Parameters for Reward Function 3.*

The mean perceived throughput performance demonstrates substantial improvement compared to the original Reward Function 3 results, achieving performance nearly comparable to the Always-ON baseline as shown in Figure 4.24 and Figure 4.25. The performance remains consistent across all three cells, indicating that the agent has successfully learned to prioritise QoS while simultaneously achieving NES gains. The throughput CDF reveals that mean perceived throughput falls to zero for less than 10% of the time, further confirming the improved network performance. This enhanced QoS is also reflected in the delay characteristics presented in Figure 4.29 and Figure 4.30, which demonstrate correspondingly improved latency behaviour.

Analysis of the predicted ON duration distributions in Figure 4.26 reveals that the modified parameter configuration produces substantially more stable behaviour across all deployment configurations. The agent selects an ON duration of 300ms (0.9375 times the cycle duration) for more than 90% of the time in each cell, demonstrating a strong preference for maintaining high capacity to ensure QoS. The percentage of time spent at ON durations exceeding 300ms decreases for sparser traffic scenarios, indicating intuitive adaptation where the agent moderately reduces transmission opportunities when traffic demands are lower. This behaviour represents a marked improvement over the previous reward function formulations, successfully balancing energy efficiency objectives with quality of service requirements.

The static power consumption characteristics in Figure 4.27 and Figure 4.28 demonstrate significant energy savings compared to the Always-ON baseline. The

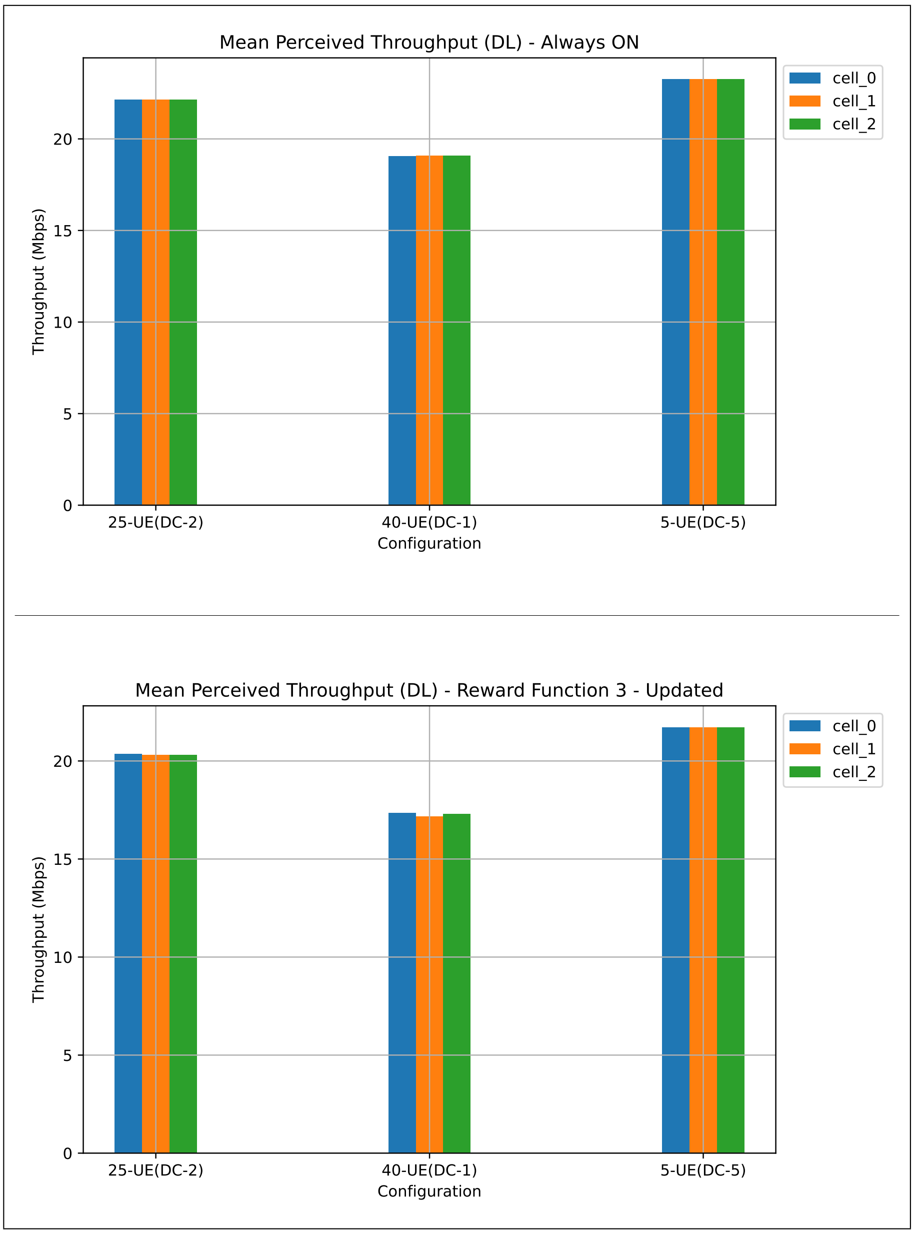

Figure 4.24 *Mean Perceived Throughput Comparison for Reward Function 3 After Parameter Update.*

CDFs show that all three cells achieve consistently lower power consumption across all deployment configurations, with the solid curves (agent performance) positioned substantially to the left of the dotted curves (Always-ON baseline). The power consumption remains remarkably consistent across different traffic intensities, with all deployment configurations clustering around similar consumption levels. This indicates that the agent maintains a stable energy-efficient operational mode while achieving near-baseline QoS performance, successfully validating the energy saving potential of the approach.

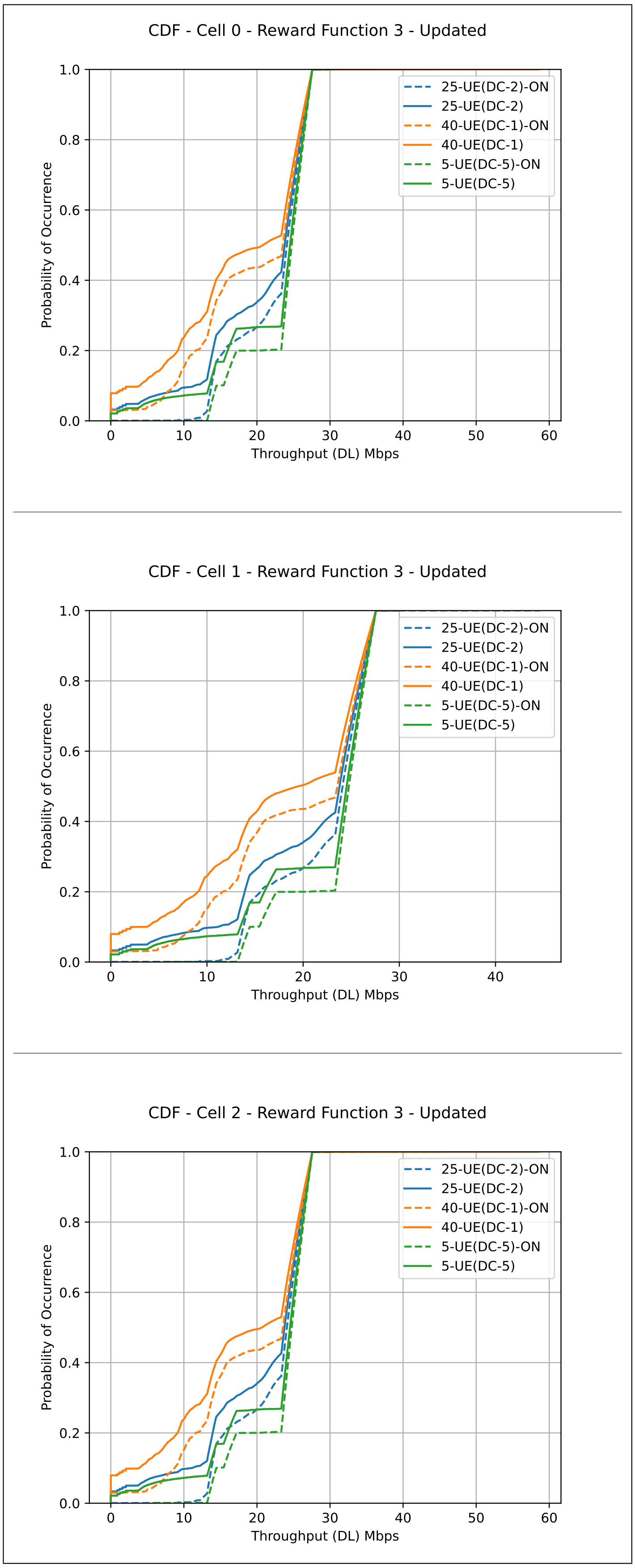

Figure 4.25 *Perceived Throughput CDFs for Reward Function 3 After Parameter Update.*

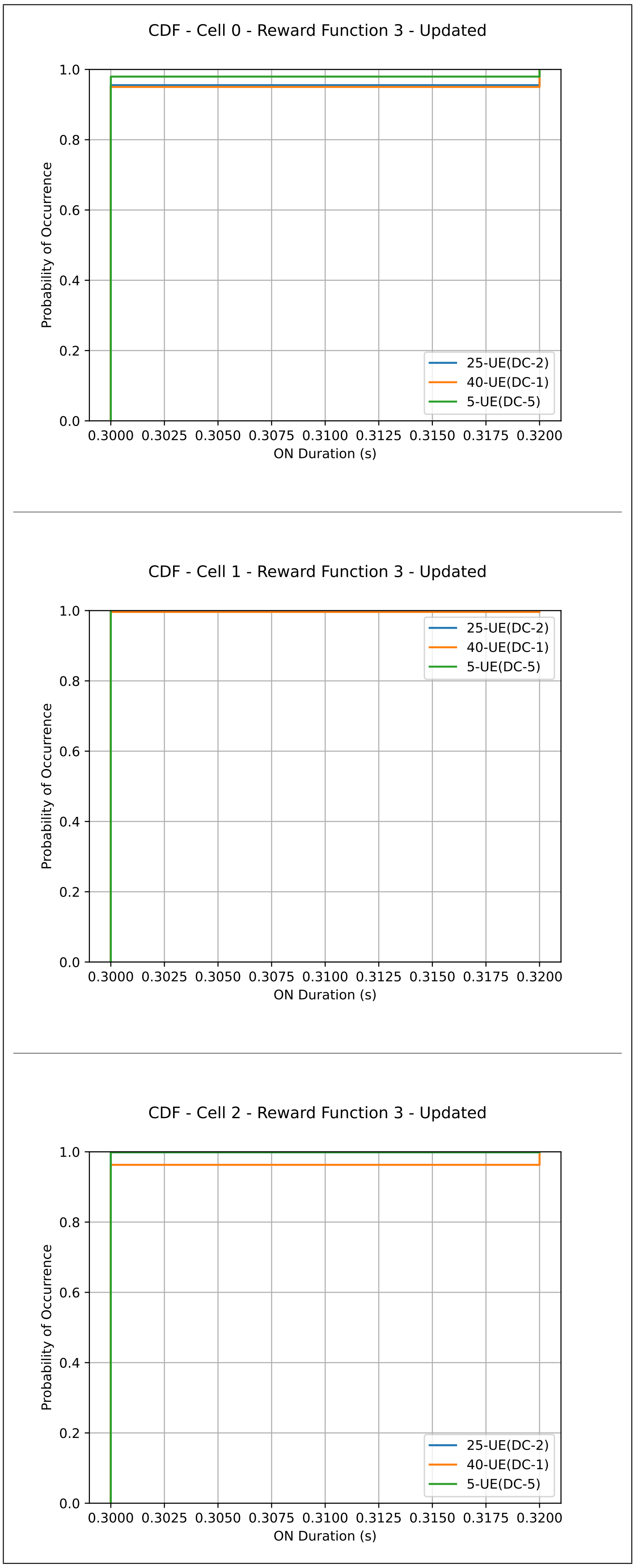

Figure 4.26 *Predicted Actions CDFs for Reward Function 3 After Parameter Update.*

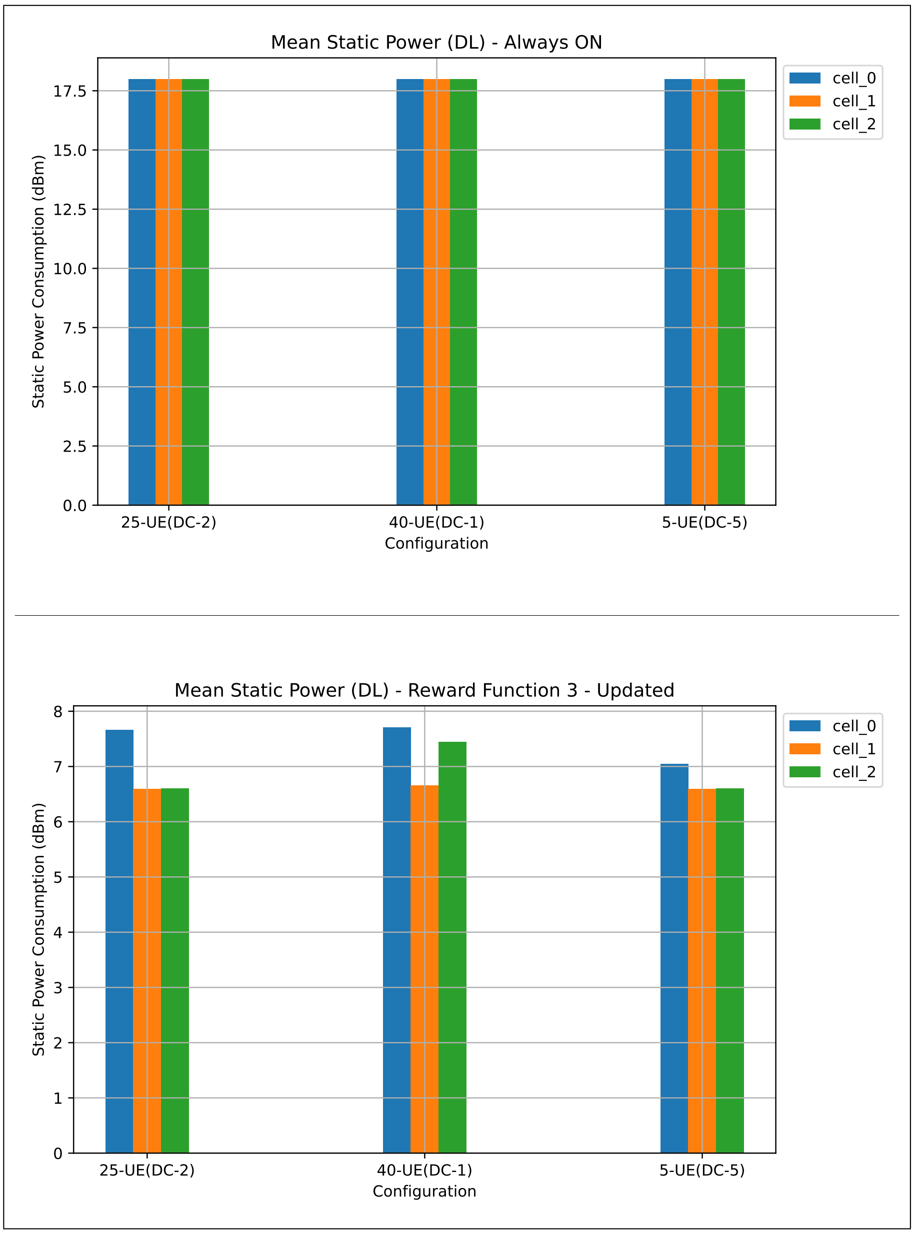

Figure 4.27 *Mean Static Power Consumption Comparison for Reward Function 3 After Parameter Update.*

This behaviour represents a marked improvement over the previous reward function formulations, successfully balancing EE objectives with QoS requirements. The modified Reward Function 3 configuration achieves substantially better performance across both energy saving and QoS dimensions compared to all previous formulations. Compared to the always-on baseline, static power consumption reductions exceeding 50% are achieved while maintaining mean throughput degradation within 10% of baseline performance across configurations spanning from dense traffic to sparse traffic scenarios. The parameter adjustments successfully shifted the agent's policy towards prioritising network performance while still achieving meaningful

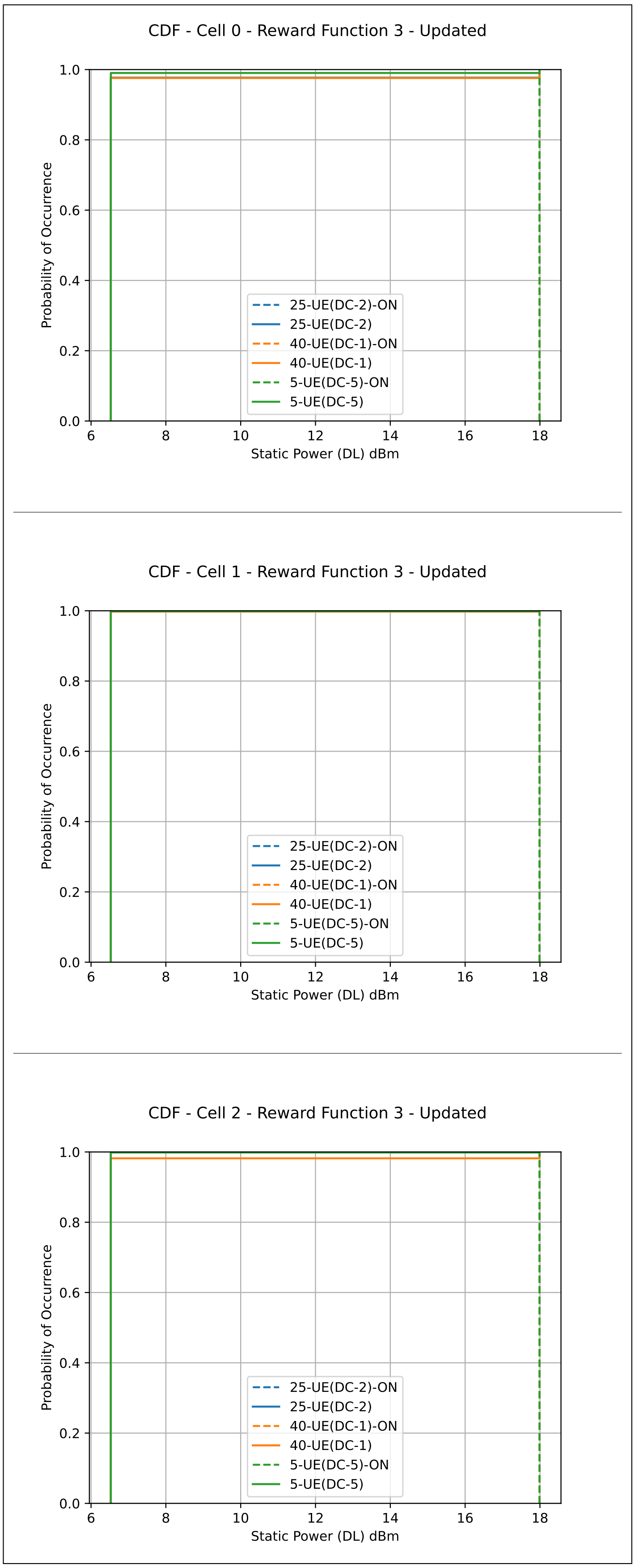

Figure 4.28 *Static Power Consumption CDFs for Reward Function 3 After Parameter Update.*

power savings relative to the Always-ON baseline. These results demonstrate that carefully designed reward functions, combined with appropriate parameter tuning, enable effective DRL-based optimization of DTX/DRX configurations in multi-cell 5G networks. The findings establish the viability of the proposed approach for adaptive energy management, validating both the simulation framework and the RL methodology presented in this thesis.

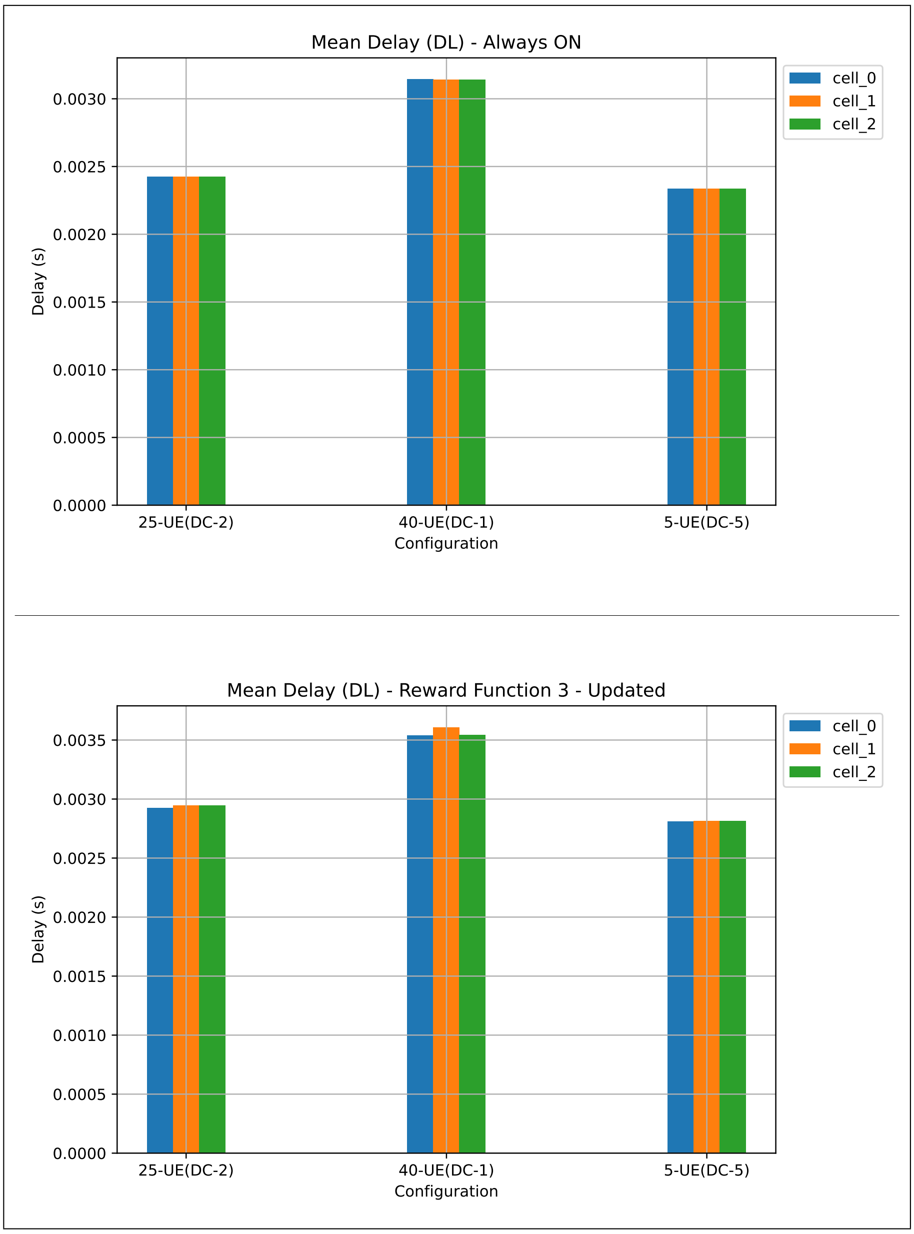

Figure 4.29 *Mean Delay Comparison for Reward Function 3 After Parameter Update.*

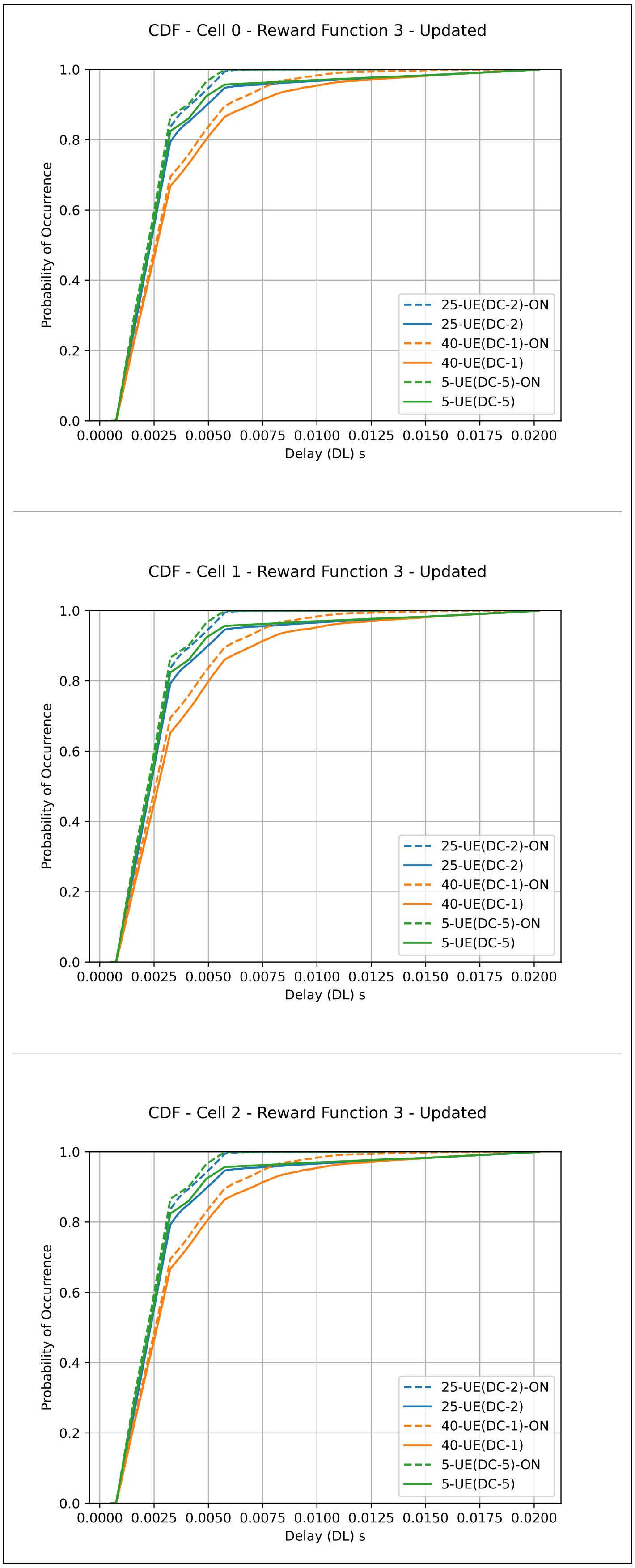

Figure 4.30 *Delay CDFs for Reward Function 3 After Parameter Update.*

5 Conclusions and Open Research Directions

5.1 Conclusions

This thesis investigated the application of DRL for energy-efficient network operation in 5G-Advanced systems, specifically focusing on the optimization of DTX/-DRX configurations in multi-cell deployments. The research addressed the challenge of balancing energy savings with QoS maintenance. The research makes several key contributions to the field of network energy saving in 5G systems.

The primary contribution of this work is the development and validation of a DRL-based framework that optimizes DTX/DRX ON duration across multiple cells in response to dynamic traffic conditions. Through systematic experimentation with three reward function formulations, the research demonstrated the critical importance of reward design in aligning agent behaviour with network optimization objectives. The initial reward functions revealed fundamental challenges including reward misalignment and null gradient problems. These findings led to the development of a more sophisticated state-dependent reward structure that successfully addresses these limitations by employing separate reward components for high-backlog and low-backlog network states.

The final reward function configuration, validated through evaluation across different deployment scenarios ranging from dense (40 UEs per cell, 80% PRB utilization) to sparse (5 UEs per cell, 10% PRB utilization) traffic conditions, achieved substantial improvements over previous formulations. The trained agent demonstrated the ability to maintain QoS performance comparable to the Always-ON baseline while achieving meaningful static power consumption reductions across diverse deployment configurations. Static power reductions surpassing 50% relative to always-on operation were obtained, with throughput maintained above 90% of baseline levels and delay characteristics remaining within acceptable service thresholds. These outcomes confirm the practical viability of the DRL-based framework for adaptive energy management in multi-cell 5G networks.

The SLS enhanced for this research provides a valuable platform for evaluating

DRL-based network optimization strategies. These enhancements included multi-cell capability, interference modelling and dynamic power consumption model. This improved simulation framework enables reliable assessment of RL frameworks without the risks and costs associated with live network experimentation.

Researchers and practitioners can leverage this work in several ways. The iterative reward function development process documented in this thesis provides practical insights into the challenges of applying RL to network optimization problems, particularly regarding reward specification and the alignment of optimization objectives with actual performance requirements. The state-dependent reward structure developed here offers a template for addressing similar multi-objective optimization challenges in other network management contexts. The SLS framework can serve as a foundation for extending the approach to incorporate additional optimization dimensions such as antenna configuration and transmission power control. Furthermore, the systematic evaluation methodology employed across diverse deployment configurations demonstrates best practices for assessing generalization capabilities of learned policies.

However, several limitations of this study should be acknowledged. First, computational constraints necessitated the use of a hexagonal 3-cell network topology for training and evaluation. While this configuration provides meaningful insights into multi-cell coordination challenges, extending to higher-order topologies such as 7-cell or 19-cell deployments would exponentially increase computational complexity. The scalability of the approach to larger network configurations remains an open question requiring further investigation. The research focused exclusively on achieving energy savings through DTX/DRX optimization impacting ASMs. Real-world deployments could potentially benefit from combining this approach with complementary techniques such as antenna muting, dynamic antenna reconfiguration, and transmission power adaptation. Such multi-dimensional optimization may improve both energy savings and solution robustness, though it would also increase the complexity of the action space and reward design. While the modified Reward Function 3 demonstrates substantial improvement over initial formulations, the sensitivity of agent behaviour to parameter adjustments suggests that the reward function may still contain subtle misalignments or sub-optimalities. State space

Despite these limitations, the research validates the fundamental viability of DRL for energy savings in 5G networks. The successful development of effective reward functions, combined with the demonstration of policy generalization across diverse traffic conditions, establishes a foundation for practical deployment of DRL-based network optimization strategies.

5.2 Open Research Directions

Several promising directions for future research emerge from this work, offering opportunities to extend the capabilities and applicability of DRL-based network energy management.

As discussed in Chapter 3, the current training methodology focuses on a single deployment configuration (DC-1 with 40 UEs per cell), with generalization assessed through evaluation on other configurations. While the trained agent demonstrates reasonable adaptation to different traffic intensities, analysis of ON duration distributions suggests conservative behaviour in sparser traffic scenarios. Future work could extend the training methodology to incorporate experiences from multiple deployment configurations simultaneously through curriculum learning or multi-task training approaches.

Moreover, the current framework employs a single centralized agent that observes the state of all cells and selects DTX/DRX configurations for the entire network. While this approach successfully demonstrates coordinated multi-cell optimization, it represents only one point in the design space of multi-agent and hierarchical RL architectures. Future research could explore decomposed approaches where multiple agents operate at different hierarchical levels or optimize different aspects of network configuration.

The current implementation employs PPO, which is an algorithm that learns expected returns under the policy. However, 5G networks exhibit significant randomness in channel conditions, interference patterns, and traffic arrivals. Distributional RL algorithms, which learn entire distributions over returns rather than only expected values, could provide improved robustness to this inherent uncertainty. Algorithms such as Quantile Regression DQN (QR-DQN) [52], Truncated Quantile Critics (TQC) [53], and Categorical DQN (C51) [54] maintain representations of return distributions, enabling risk-sensitive decision making. In the network energy management context, this could allow the agent to distinguish between actions with similar expected performance but different risk profiles. A distributional RL agent could learn to recognize such trade-offs and adopt accordingly.

The current work focuses exclusively on DTX/DRX configuration as the mechanism for energy saving via ASMs. However, operational 5G networks employ multiple complementary techniques for energy management, including antenna muting, dynamic spectrum management, and transmission power control. Future research could extend the DRL framework to jointly optimize across these multiple dimensions. However, such multi-dimensional optimization presents both opportunities and challenges.

These research directions collectively offer pathways to extend the capabilities, robustness, and practical applicability of DRL for 5G network energy management. Addressing these challenges would advance the field toward deployment-ready solutions that can meaningfully contribute to sustainable network operation.

References

[1] IEA. *Energy Efficiency 2024*. Report. Nov. 2024. URL: https://www.iea.org/reports/energy-efficiency-2024.

[2] IEA. *CO2 Emissions in 2023 : A new record high, but is there light at the end of the tunnel?* Report. Mar. 2024. URL: https://www.iea.org/reports/co2-emissions-in-2023.

[3] ITU. *Population of global offline continues steady decline to 2.6 billion people in 2023*. Press Release. Sept. 2023. URL: https://www.itu.int/en/mediacentre/Pages/PR-2023-09-12-universal-and-meaningful-connectivity-by-2030.aspx.

[4] ITU/WB. *Measuring the Emissions & Energy Footprint of the ICT Sector*. Report. 2024. URL: https://documents1.worldbank.org/curated/en/099121223165540890/pdf/P17859712a98880541a4b71d57876048abb.pdf.

[5] Hannah Ritchie. "What share of global CO emissions come from aviation?" In: *Our World in Data* (2024). URL: https://ourworldindata.org/global-aviation-emissions.

[6] Ericsson. *Ericsson Mobility Report*. Nov. 2024. URL: https://www.ericsson.com/en/reports-and-papers/mobility-report.

[7] Nokia. *5G Network Energy Efficiency – Nokia White Paper*. White Paper. Jan. 2017. URL: https://gsacom.com/paper/5g-network-energy-efficiency-nokia-white-paper/.

[8] David López-Pérez et al. "A Survey on 5G Radio Access Network Energy Efficiency: Massive MIMO, Lean Carrier Design, Sleep Modes, and Machine Learning". In: *IEEE Communications Surveys & Tutorials* 24.1 (2022), pp. 653–697. DOI: 10.1109/COMST.2022.3142532.

[9] 3GPP. *Releases*. Specifications and Technologies. 2024. URL: https://www.3gpp.org/specifications-technologies/releases/.

[10] Fatma Ezzahra Salem et al. "Advanced Sleep Modes and Their Impact on Flow-Level Performance of 5G Networks". In: *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. 2017, pp. 1–7. DOI: 10.1109/VTCFall.2017.8288125.

[11] Fatma Ezzahra Salem et al. "Traffic-aware Advanced Sleep Modes management in 5G networks". In: *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. 2019, pp. 1–6. DOI: 10.1109/WCNC.2019.8886051.

[12] Rumeng Tan et al. "Energy Saving Technologies and Best Practices for 5G Radio Access Network". In: *IEEE Access* 10 (2022), pp. 51747–51756. DOI: 10.1109/ACCESS.2022.3174089.

[13] Sandra Lagen et al. "New Radio Physical Layer Abstraction for System-Level Simulations of 5G Networks". In: *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. 2020, pp. 1–7. DOI: 10.1109/ICC40277.2020.9149444.

[14] Chin-Kuo Jao et al. "WiSE: A System-Level Simulator for 5G Mobile Networks". In: *IEEE Wireless Communications* 25.2 (2018), pp. 4–7. DOI: 10.1109/MWC.2018.8352614.

[15] Conrado Boeira et al. "Calibration and Automation of a 5G Simulator for Realistic Evaluation and Data Generation". In: *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*. 2024, pp. 341–345. DOI: 10.1109/NetSoft60951.2024.10588948.

[16] Miguel Cantero et al. "System-Level Performance Evaluation of 5G Use Cases for Industrial Scenarios". In: *IEEE Access* 11 (2023), pp. 37778–37789. DOI: 10.1109/ACCESS.2023.3266981.

[17] Dedy Irawan et al. "Network Slicing Modeling and Simulation on 5G Networks Using Virne". In: *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. 2023, pp. 114–119. DOI: 10.1109/IC3INA60834.2023.10285790.

[18] Thilina Weththasinghe, Marja Matinmikko-Blue, and Arturo Basaure. "Simulation of Interference in 5G/6G Networks Operating in Locally Licensed Spectrum". In: *2025 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*. 2025, pp. 1–10. DOI: 10.1109/DySPAN64764.2025.11115930.

[19] Nokia. *Reducing energy use with 5G-Advanced – Nokia White Paper*. Nov. 2023. URL: https://www.nokia.com/asset/f/213599/.

[20] Wei Mao et al. *Deep Reinforcement Learning-based Cell DTX/DRX Configuration for Network Energy Saving*. July 2025. DOI: 10.48550/arXiv.2507.21385.

[21] Ali El-Amine et al. "A Distributed Q-Learning Approach for Adaptive Sleep Modes in 5G Networks". In: *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. 2019, pp. 1–6. DOI: 10.1109/WCNC.2019.8885818.

[22] Fatma Ezzahra Salem et al. "Reinforcement Learning Approach for Advanced Sleep Modes Management in 5G Networks". In: *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*. 2018, pp. 1–5. DOI: 10.1109/VTCFall.2018.8690555.

[23] Yusuf Usman et al. "AI, ML, and LLM Integration in 5G/6G Networks: A Comprehensive Survey of Architectures, Challenges, and Future Directions". In: *IEEE Access* 13 (2025), pp. 168914–168950. DOI: 10.1109/ACCESS.2025.3608736.

[24] Mustafa Al-Khafaji and Lina Elwiya. "ML/AI Empowered 5G and beyond Networks". In: *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. 2022, pp. 1–6. DOI: 10.1109/HORA55278.2022.9799813.

[25] Aqsa Sayed. "Intelligent Optimization of LTE and 5G Networks Using AI and Machine Learning". In: *International Journal of Artificial Intelligence, Data Science, and Machine Learning* 6.3 (Oct. 2025), pp. 24–31. DOI: 10.63282/3050-9262.IJAIDSML-V6I3P104. URL: https://ijaidsml.org/index.php/ijaidsml/article/view/233.

[26] Manish Kumar. "AI/ML Driven 5G Networks: From Deployment to Optimization". In: *International Journal of Innovative Science and Research Technology* (Sept. 2025), pp. 162–168. DOI: 10.38124/ijisrt/25sep051.

[27] Matteo Bordin et al. "Design and Evaluation of Deep Reinforcement Learning for Energy Saving in Open RAN". In: *2025 IEEE 22nd Consumer Communications Networking Conference (CCNC)*. 2025, pp. 1–6. DOI: 10.1109/CCNC54725.2025.10976108.

[28] Hasna Fourati et al. "An efficient energy saving scheme using reinforcement learning for 5G and beyond in H-CRAN". In: *Ad Hoc Networks* 155 (2024), p. 103406. ISSN: 1570-8705. DOI: https://doi.org/10.1016/j.adhoc.2024.103406. URL: https://www.sciencedirect.com/science/article/pii/S1570870524000179.

[29] Sami Alenezi, Chunbo Luo, and Geyong Min. "Energy-Efficient D2D Communications Based on Centralised Reinforcement Learning Techniques". In: *2021 IEEE 24th International Conference on Computational Science and Engineering (CSE)*. 2021, pp. 57–63. DOI: 10.1109/CSE53436.2021.00018.

[30] Yaser Azimi et al. "Energy-Efficient Deep Reinforcement Learning Assisted Resource Allocation for 5G-RAN Slicing". In: *IEEE Transactions on Vehicular Technology* 71.1 (2022), pp. 856–871. DOI: 10.1109/TVT.2021.3128513.

[31] Vodafone and Ericsson. *Vodafone and Ericsson trial 5G energy efficiency solutions.* Accessed: 2025-11-03. Mar. 2025. URL: https://www.ericsson.com/en/press-releases/3/2025/vodafone-uk-and-ericsson-trial-ai-solutions-for-improved-5g-energy-efficiency.

[32] 3GPP TR 38.743. *Study on enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN.* Sept. 2024. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4286.

[33] Toufiqul Islam, Daewon Lee, and Seau Sian Lim. "Enabling Network Power Savings in 5G-Advanced and Beyond". In: *IEEE Journal on Selected Areas in Communications* 41.6 (2023), pp. 1888–1899. DOI: 10.1109/JSAC.2023.3273706.

[34] 3GPP TR 38.864. *Study on Network Energy Savings for NR.* Dec. 2022. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3987.

[35] Yu-Ngok Ruyue Li et al. "Power Saving Techniques for 5G and Beyond". In: *IEEE Access* 8 (2020), pp. 108675–108690. DOI: 10.1109/ACCESS.2020.3001180.

[36] Bjorn Debaillie, Claude Desset, and Filip Louagie. "A Flexible and Future-Proof Power Model for Cellular Base Stations". In: *2015 IEEE 81st Vehicular Technology Conference (VTC Spring).* 2015, pp. 1–7. DOI: 10.1109/VTCSpring.2015.7145603.

[37] Fatma Ezzahra Salem et al. "Advanced Sleep Modes and Their Impact on Flow-Level Performance of 5G Networks". In: *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall).* 2017, pp. 1–7. DOI: 10.1109/VTCFall.2017.8288125.

[38] 3GPP TS 38.331. *NR; Radio Resource Control (RRC); Protocol specification.* Dec. 2024. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197.

[39] 3GPP TR 37.817. *Study on enhancement for data collection for NR and ENDC.* Apr. 2022. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3817.

[40] 3GPP TR 22.882. *Study on Energy Efficiency as service criteria.* Apr. 2024. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4096.

[41] Nokia. *AI/ML for NG-RAN*. June 2023. URL: https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_AHs/2023_06_RAN_Rel19_WS/Docs/RWS-230023.zip.

[42] ETSI ES 203 228 V1.4.1. *Environmental Engineering (EE); Assessment of mobile network energy efficienc*. Apr. 2022. URL: https://www.etsi.org/deliver/etsi_es/203200_203299/203228/01.04.01_60/es_203228v010401p.pdf.

[43] 3GPP TS 28.554. *Management and Orchestration; 5G end to end Key Performance Indicators (KPI)*. Dec. 2023. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3056.

[44] O-RAN.WG7.NES.0-v02.00. *Network Energy Savings Procedures and Performance Metrics*. Oct. 2024. URL: https://specifications.o-ran.org/download?id=758.

[45] 3GPP TR 38.801. *Study on new radio access technology: Radio access architecture and interfaces*. Apr. 2017. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3056.

[46] Yekaterina Sadovaya et al. "Delay-Aware Link Scheduling in IAB Networks With Dynamic User Demands". In: *IEEE Transactions on Vehicular Technology* 73.10 (2024), pp. 15125–15139. DOI: 10.1109/TVT.2024.3409179.

[47] Yekaterina Sadovaya et al. "Impact of System-Specific Factors on Scheduling and Resource Allocation in mmWave IAB Networks". In: *ICC 2024 - IEEE International Conference on Communications*. 2024, pp. 43–48. DOI: 10.1109/ICC51166.2024.10622890.

[48] Yekaterina Sadovaya et al. "Integrated Access and Backhaul in Millimeter-Wave Cellular: Benefits and Challenges". In: *IEEE Communications Magazine* 60.9 (2022), pp. 81–86. DOI: 10.1109/MCOM.004.2101082.

[49] IEEE 802.16 Broadband Wireless Access Working Group. *IEEE 802.16m Evaluation Methodology Document*. Tech. rep. IEEE 802.16m-08/004r5. EMD. Institute of Electrical and Electronics Engineers, Jan. 2009.

[50] Antonin Raffin et al. *Stable-Baselines3*. https://github.com/DLR-RM/stable-baselines3. Version 2.4.0. 2024.

[51] TensorFlow Developers. *TensorBoard: TensorFlow's Visualization Toolkit*. https://www.tensorflow.org/tensorboard.

[52] Will Dabney et al. "Distributional reinforcement learning with quantile regression". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (2018). DOI: 10.1609/aaai.v32i1.11791.

[53] Arsenii Kuznetsov et al. "Controlling overestimation bias with truncated mixture of continuous distributional quantile critics". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5556–5566.

[54] Marc G Bellemare, Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 449–458.