

D. Marlambekov^{1*}, S. Mukhambetzhanov¹, S. Ibraheem²¹Al-Farabi Kazakh National University, Almaty, Kazakhstan²Istanbul Technical University, Istanbul, Turkey*e-mail: marlambekov_duman@live.kaznu.kz

PPO-BASED RESOURCE ALLOCATION FOR FIFTH-GENERATION NETWORK SLICING

Abstract

This paper presents a deep reinforcement learning approach to service-level-agreement (SLA) aware network slicing in 5G radio access networks. We formulate the dynamic allocation of Physical Resource Blocks (PRB) among three coexisting slices – enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC) – as a Markov Decision Process and develop a Gymnasium-compatible simulation environment in which per-slice SLA requirements follow the 3GPP TS 28.530 framework. A Proximal Policy Optimization (PPO) agent is trained with a composite reward that jointly rewards aggregate throughput, penalizes SLA violations, and encourages resource utilization. The agent is evaluated over 100 episodes against three classical baselines: Static Equal, Static Priority, and Greedy Proportional allocation. The PPO agent reduces the aggregate SLA violation rate to 2×10^{-5} , three to four orders of magnitude lower than the Static Priority baseline (0.46), while sustaining a mean aggregate throughput of 74.7 Mbit/s and a resource utilization of 98.6%. A throughput-versus-violation analysis confirms that the learned policy occupies an operating point that no single static rule attains. The complete environment, training pipeline, and evaluation scripts are open-source and reproducible using the Stable-Baselines3 library.

Keywords: 5G network slicing, reinforcement learning, proximal policy optimization, service-level agreement, PRB allocation, eMBB, URLLC, mMTC, radio resource management.

Д. Марламбеков¹, С. Мұхамбетжанов¹, Ш. Ибрахим²¹ Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан² Стамбул техникалық университеті, Стамбул қ., Түркия

PPO НЕГІЗІНДЕ БЕСІНШІ БУЫН ЖЕЛІЛІК СЛАЙСИНГІ ҮШІН РЕСУРСТАРДЫ БӨЛУ

Аңдатпа

Бұл жұмыста 5G радиоқолжетімділік желілеріндегі қызмет көрсету деңгейі туралы келісімдерді (SLA) ескеретін желілік слайсингке арналған тереңдетілген оқыту тәсілі ұсынылған. Үш қатар жұмыс істейтін слайс – кеңейтілген мобильді кең жолақты байланыс (eMBB), аса сенімді және кідірісі төмен байланыс (URLLC) және жаппай машиналық типтегі байланыс (mMTC) – арасында физикалық ресурстық блоктарды (PRB) динамикалық бөлу міндеті Марков шешім қабылдау процесі ретінде тұжырымдалған. Әрбір слайсқа қойылатын SLA талаптары 3GPP TS 28.530 стандартына сәйкес келетін, Gymnasium ортасымен үйлесімді имитациялық орта әзірленді. Proximal Policy Optimization (PPO) агенті жиынтық өткізу қабілетін ынталандыратын, SLA бұзушылықтарын айыптайтын және ресурстарды тиімді пайдалануды көтермелейтін құрамдас марапат функциясы арқылы оқытылады. Агент 100 эпизод бойынша бағаланып, Static Equal, Static Priority және Greedy Proportional атты үш классикалық базалық стратегиямен салыстырылды. PPO агенті SLA бұзушылықтарының жиынтық жиілігін 2×10^{-5} деңгейіне дейін төмендетті, бұл Static Priority базалық стратегиясымен салыстырғанда (0,46) үш-төрт реттік шамаға аз. Сонымен қатар орташа жиынтық өткізу қабілеті 74,7 Мбит/с, ал ресурстарды пайдалану деңгейі 98,6% болды. Өткізу қабілеті мен бұзушылықтар арасындағы ымыраны талдау үйренген саясаттың бірде-бір статикалық стратегия қол жеткізе алмайтын жұмыс нүктесіне жететінін көрсетті. Толық имитациялық орта, оқыту үдерісі және бағалау скрипттері ашық түрде ұсынылған және Stable-Baselines3 кітапханасы арқылы қайта өндіріледі.

Түйін сөздер: 5G желілік слайсингі, күшейтпелі оқыту, Proximal Policy Optimization, қызмет көрсету деңгейі туралы келісім, PRB бөлу, eMBB, URLLC, mMTC, радиоресурстарды басқару.

Д. Марламбеков^{1*}, С. Мухамбетжанов¹, Ш. Ибрахим²

¹Казахский национальный университет имени аль-Фараби, г.Алматы, Казахстан

²Стамбульский технический университет, г.Стамбул, Турция

РАСПРЕДЕЛЕНИЕ РЕСУРСОВ НА ОСНОВЕ РРО ДЛЯ СЕГМЕНТАЦИИ СЕТЕЙ ПЯТОГО ПОКОЛЕНИЯ

Аннотация

В данной работе представлен подход на основе глубокого обучения с подкреплением к сетевому слайсингу в сетях радиодоступа 5G с учётом соглашений об уровне обслуживания (SLA). Задача динамического распределения физических ресурсных блоков (PRB) между тремя одновременно функционирующими слайсами – усовершенствованной мобильной широкополосной связью (eMBB), сверхнадёжной связью с малой задержкой (URLLC) и массовой межмашинной связью (mMTC) – формулируется как марковский процесс принятия решений. Разработана совместимая с Gymnasium имитационная среда, в которой требования SLA для каждого слайса соответствуют стандарту 3GPP TS 28.530. Агент Proximal Policy Optimization (PPO) обучается с использованием составной функции вознаграждения, которая поощряет совокупную пропускную способность, штрафует нарушения SLA и стимулирует эффективное использование ресурсов. Агент оценивается в течение 100 эпизодов и сравнивается с тремя классическими базовыми стратегиями: Static Equal, Static Priority и Greedy Proportional. Агент PPO снижает совокупную частоту нарушений SLA до 2×10^{-5} , что на три-четыре порядка ниже по сравнению с базовой стратегией Static Priority (0,46), при этом обеспечивая среднюю совокупную пропускную способность 74,7 Мбит/с и коэффициент использования ресурсов 98,6%. Анализ компромисса между пропускной способностью и частотой нарушений показывает, что обученная политика достигает рабочей точки, недоступной ни одной отдельной статической стратегии. Полная имитационная среда, процесс обучения и скрипты оценки являются открытыми и воспроизводимыми с использованием библиотеки Stable-Baselines3.

Ключевые слова: сетевой слайсинг 5G, обучение с подкреплением, Proximal Policy Optimization, соглашение об уровне обслуживания, распределение PRB, eMBB, URLLC, mMTC, управление радиоресурсами.

Introduction

Main Provisions

A Gymnasium-compatible simulation environment was developed for SLA-aware PRB allocation across eMBB, URLLC, and mMTC slices, with per-slice contracts aligned to the 3GPP TS 28.530 management framework. A PPO agent trained with a composite throughput-SLA-utilization reward was benchmarked over 100 episodes against Static Equal, Static Priority, and Greedy Proportional baselines. The learned policy reduced the aggregate SLA violation rate to 2×10^{-5} – three to four orders of magnitude below the Static Priority baseline – while sustaining 74.7 Mbit/s of throughput and 98.6% resource utilization, occupying an operating point on the throughput-violation frontier that no static rule reaches. These results establish reinforcement learning with an explicit, decomposed SLA objective as a practical path toward adaptive, contract-compliant network slice management.

Network slicing is one of the defining capabilities of fifth-generation (5G) mobile systems. Mobile data traffic continues to grow and is carried by increasingly diverse services with very different performance requirements [1, 2]. Slicing allows a single physical infrastructure to be partitioned into multiple logical networks, each tailored to a distinct class of service with its own performance contract [3, 4]. The three canonical slice types standardized for 5G – enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC) – impose fundamentally different and often conflicting demands on the radio scheduler [5, 6]. eMBB traffic seeks high sustained data rates, URLLC traffic requires stringent reliability and deadline guarantees, and mMTC traffic must connect a very large number of low-rate devices. Allocating the finite pool of Physical Resource Blocks (PRB) at the base station so that every slice meets its service-level agreement (SLA) is therefore a continuous, multi-objective decision problem under uncertainty.

Traditional slicing controllers rely on fixed allocation policies. Equal partitioning divides PRBs uniformly across slices, static priority schemes reserve a guaranteed share for the most critical slice, and traffic-proportional rules distribute resources according to instantaneous load. While simple and predictable, these heuristics cannot adapt to the non-stationary statistics of mixed 5G traffic: a policy tuned for one load regime systematically violates SLAs in another [7, 8]. The combinatorial structure of the allocation, together with the stochastic arrival of eMBB, URLLC, and mMTC demand, makes closed-form optimal control intractable for online deployment [9].

Deep reinforcement learning (DRL) has emerged as a powerful data-driven alternative for radio resource management. Following the demonstration that neural networks can learn control policies directly from interaction [10], a substantial body of work has applied DRL to slicing. Li et al. introduced DRL for resource management across slices [7], Raza et al. applied reinforcement learning to slice admission in a flexible 5G RAN [11], and distributional and generative extensions further improved sample efficiency [12]. Recent studies address end-to-end slicing with multi-agent actor-critic methods [13], slice scaling and placement [14], admission control [15, 16], and constrained formulations that embed SLA limits directly into the learning objective [17]. Comprehensive surveys confirm that learned policies consistently dominate fixed heuristics across heterogeneous network conditions [8, 14]. Among policy-gradient methods, Proximal Policy Optimization (PPO) [18] has become the algorithm of choice for slicing control because its clipped surrogate objective yields stable training over the large action spaces typical of PRB allocation, as demonstrated in O-RAN xApp deployments [19] and reward-shaped orchestration frameworks [20].

Despite this progress, several gaps persist. First, many DRL slicing studies optimize an abstract utility and do not verify performance against explicit, standardized SLA contracts, making the operational relevance of reported gains difficult to assess [8, 21]. Second, the per-slice nature of SLA satisfaction – distinguishing a throughput floor for eMBB, a reliability target for URLLC, and a served-device ratio for mMTC – is frequently collapsed into a single scalar, obscuring where violations actually occur. Third, systematic benchmarking against multiple interpretable static baselines within one reproducible environment remains uncommon [22]. The present work targets these gaps. The central research question is: can a PPO agent learn a slicing policy that simultaneously keeps the per-slice SLA violation rate near zero and sustains high aggregate throughput and resource utilization, outperforming Static Equal, Static Priority, and Greedy Proportional allocation within a 3GPP-aligned simulation environment? Our contributions are fourfold. (i) We develop a Gymnasium-compatible single-cell slicing environment in which eMBB, URLLC, and mMTC SLA requirements follow the 3GPP TS 28.530 management framework [3], with a Shannon-based throughput model driven by CQI-to-spectral-efficiency mapping. (ii) We design a PPO agent with a composite reward that jointly rewards normalized throughput, penalizes aggregate SLA violation, and rewards utilization. (iii) We benchmark the agent against three deterministic baselines over 100 evaluation episodes, reporting per-slice violation rates and PRB distributions. (iv) We show through a throughput-versus-violation trade-off analysis that the learned policy reaches an operating region inaccessible to any single static rule. All code and results are open-source and reproducible with Stable-Baselines3 [23].

Research methodology

Network and Traffic Model

We consider the downlink of a single 5G base station that serves three slices simultaneously. The cell owns a fixed pool of $N_{\text{PRB}} = 100$ Physical Resource Blocks, each spanning $B_{\text{sc}} = 180$ kHz (twelve 15-kHz subcarriers). Decisions are taken once per Transmission Time Interval (TTI); an episode spans 500 TTIs. At each TTI the traffic model draws the instantaneous demand of every slice from independent Poisson processes: the number of active eMBB, URLLC, and mMTC users follows Poisson distributions with means of 20, 5, and 50, respectively, reflecting a broadband-moderate, critical-sparse, and massive-but-light demand profile. Each user is assigned a Channel Quality

Indicator (CQI) drawn from a slice-specific range – high for URLLC, moderate for eMBB, and low for mMTC – capturing the typical propagation conditions of each service class.

The spectral efficiency of a user with channel quality index q is obtained from the 3GPP CQI-to-SNR mapping [24] followed by the Shannon bound as shown in Equation 1:

$$SE(q) = \log_2(1 + 10^{(SNR(q)/10)}) \quad (1)$$

where $SNR(q)$ is the signal-to-noise ratio tabulated for CQI index q . The throughput (Equation 2) delivered to a slice that receives k PRBs whose users have mean spectral efficiency $\bar{SE}$ is then:

$$R = k \cdot B_{sc} \cdot \bar{SE} \quad [bit/s] \quad (2)$$

where $B_{sc} = 180$ kHz is the PRB bandwidth. The theoretical cell capacity used to normalize throughput in the reward, denoted R_{max} , is obtained by assigning all PRBs at the maximum CQI across the three slices.

Per-Slice SLA Requirements (3GPP TS 28.530)

Each slice carries an SLA contract expressed in the terms of the 3GPP TS 28.530 management framework [3]. The eMBB slice requires a minimum throughput; the URLLC slice requires a minimum reliability under a one-TTI deadline; the mMTC slice requires a minimum fraction of served devices. Formally, for the eMBB slice the violation indicator is shown in Equation 3:

$$v_{eMBB} = \mathbb{1}[R_{eMBB} < R_{min}], \quad R_{min} = 10 \text{ Mbit/s} \quad (3)$$

For the URLLC slice, let n_p be the number of critical packets generated in the current TTI and s the number served within the deadline (Equation 4). The reliability and its violation indicator are:

$$\rho = 1 - \frac{\max(0, n_p - s)}{n_p}, \quad v_{URLLC} = \mathbb{1}[\rho < 0.999] \quad (4)$$

For the mMTC slice, where each PRB serves on average two low-rate devices, the served-device ratio (Equation 5) and its violation indicator are:

$$\sigma = \min\left(1, \frac{2 \cdot k_{mMTC}}{n_{mMTC}}\right), \quad v_{mMTC} = \mathbb{1}[\sigma < 0.80] \quad (5)$$

MDP Formulation

The slicing-control task is cast as a Markov Decision Process (S, A, P, R, γ). The state observed by the agent at each TTI is a 12-dimensional normalized vector which is shown in Equation 6:

$$s = (\ell_{eMBB}, \ell_{URLLC}, \ell_{mMTC}, k', v', \bar{q}_{eMBB}, \bar{q}_{URLLC}, \bar{q}_{mMTC}) \quad (6)$$

comprising the normalized per-slice load ℓ , the previous PRB allocation k' and previous violation flags v' (three components each), and the mean CQI of every slice. The action is a vector of requested PRB counts, one per slice (Equation 7):

$$a = (k_{eMBB}, k_{URLLC}, k_{mMTC}) \in \{0, \dots, 100\}^3 \quad (7)$$

represented as a MultiDiscrete space and rescaled so that the total never exceeds $N_{PRB} = 100$. The composite reward as shown in Equation 8, couples the three design goals – throughput, SLA compliance, and utilization:

$$r = 0.4 \cdot \frac{R_{tot}}{R_{max}} - 0.4 \cdot \frac{v_{eMBB} + v_{URLLC} + v_{mMTC}}{3} + 0.2 \cdot \frac{k_{eMBB} + k_{URLLC} + k_{mMTC}}{N_{PRB}} \quad (8)$$

where R_{tot} is the aggregate delivered throughput and R_{max} the normalizing capacity. The discount factor is $\gamma = 0.99$. This reward deliberately balances a reward term (throughput), a penalty term (SLA violation), and a regularizer (utilization), so that the agent is driven toward high-rate yet contract-compliant allocations rather than extreme ones.

PPO Agent and Baseline Policies

The control policy is learned with Proximal Policy Optimization [18], an on-policy actor-critic algorithm whose clipped surrogate objective bounds the policy update at each step and thereby stabilizes training over the large MultiDiscrete action space of Equation (7). The policy and value networks are multilayer perceptrons with two hidden layers of 64 units and tanh activations. The agent is implemented with Stable-Baselines3 [23]. Table 1 lists the hyperparameters.

We compare the learned policy against three deterministic baselines. Static Equal partitions the 100 PRBs as evenly as possible across the three slices (33/33/34) at every TTI, independent of load or channel. Static Priority reserves a fixed, generous share for the throughput- and reliability-critical slices (eMBB and URLLC) and assigns the remainder to mMTC, reflecting an operator-priority rule. Greedy Proportional distributes PRBs in proportion to the instantaneous per-slice load, a demand-following heuristic.

Table 1. PPO hyperparameters

<i>Hyperparameter</i>	<i>Value</i>
<i>Policy network</i>	<i>MLP, two layers of 64 units, tanh</i>
<i>Total timesteps</i>	<i>1,000,000</i>
<i>Parallel environments</i>	<i>4</i>
<i>Learning rate</i>	<i>3×10^{-4}</i>
<i>Rollout length (n steps)</i>	<i>2048</i>
<i>Mini-batch size</i>	<i>256</i>
<i>Entropy coefficient</i>	<i>0.01</i>
<i>Discount factor (γ)</i>	<i>0.99</i>
<i>Random seeds</i>	<i>0, 1, 2</i>

Experimental Setup

The PPO agent is trained for one million environment steps with four parallel environments and three independent random seeds (0–2) to quantify run-to-run variability. After training, every method – the PPO agent and the three baselines – is evaluated over 100 independent episodes of 500 TTIs each. For each method we record the mean aggregate throughput, the aggregate and per-slice SLA violation rates, the resource utilization, and the mean PRB allocation per slice. The environment, traffic model, SLA checker, training loop, and evaluation pipeline are released as open-source code; all reported numbers are reproducible from the accompanying repository.

Results of the Study

Figure 1 shows the PPO training reward as a function of timesteps, averaged over the three seeds with a 95% confidence band. The agent improves rapidly during the first 100,000 steps as it learns to read the per-slice load and CQI components of the observation, and then settles onto a stable plateau. The final mean reward across seeds is 134.9 ± 0.3 , and the narrow confidence band throughout the steady-state phase indicates that learning is robust to the stochastic traffic and channel draws.

Table 2 summarizes the performance of all four methods over the 100 evaluation episodes. The PPO agent drives the aggregate SLA violation rate down to 2×10^{-5} while sustaining 74.7 Mbit/s of throughput and 98.6% utilization. The Static Priority baseline attains the highest raw throughput (79.5 Mbit/s) but at a catastrophic SLA violation rate of 0.46, because its rigid reservation starves the mMTC slice whenever device counts are high. Static Equal achieves zero violations but leaves throughput on the table (69.8 Mbit/s), since it cannot shift resources toward the slices that need them. Greedy Proportional follows demand and keeps violations low (5.3×10^{-3}) but still trails the learned policy on both throughput and reliability.

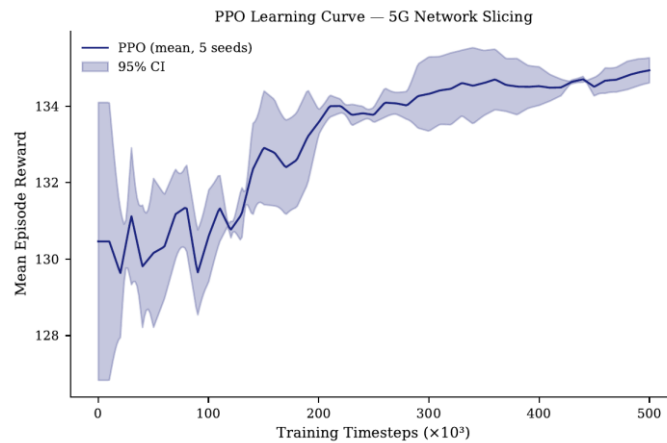


Figure 1. PPO training reward versus timesteps, averaged over three seeds (shaded area: 95% confidence interval)

Table 2. Aggregate performance over 100 evaluation episodes

Method	Throughput (Mbit/s)	SLA violation rate	Utilization	PRB eMBB/URLLC/mMTC
PPO (ours)	74.7	2×10^{-5}	0.986	28.9 / 42.2 / 27.6
Static Equal	69.8	0.000	1.000	33.0 / 33.0 / 34.0
Static Priority	79.5	0.464	1.000	40.0 / 40.0 / 20.0
Greedy Proportional	66.8	5.3×10^{-3}	1.000	35.2 / 27.8 / 37.0

Figure 2 breaks the violation rate down by slice. The static rules concentrate their failures in a single slice – Static Priority on mMTC, Greedy Proportional spread thinly across eMBB and URLLC – whereas the PPO agent keeps all three per-slice rates at or near zero, with only a negligible residual on mMTC. This is the practical meaning of SLA-aware control: not a lower average, but the elimination of the systematic per-slice failures that fixed rules incur.

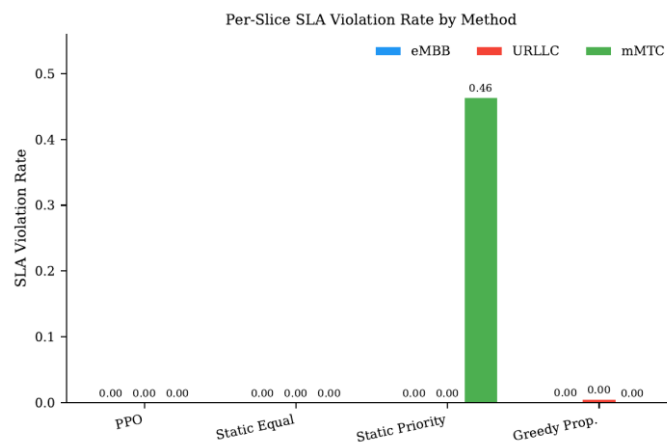


Figure 2. Per-slice SLA violation rate by method

Figure 3 plots throughput against the aggregate SLA violation rate, exposing the trade-off frontier directly. Static Priority sits far to the right (high throughput, high violation); Static Equal sits at the bottom (zero violation, modest throughput). The PPO agent occupies the desirable upper-left region – high throughput at near-zero violation – an operating point that none of the static policies can reach, confirming that the learned policy is not merely interpolating between baselines but discovering a qualitatively better allocation strategy.

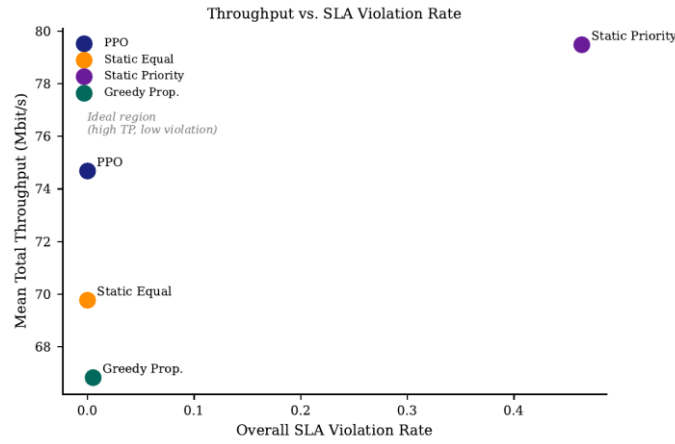


Figure 3. Throughput versus aggregate SLA violation trade-off across methods.

Figure 4 shows the mean PRB distribution chosen by each method.

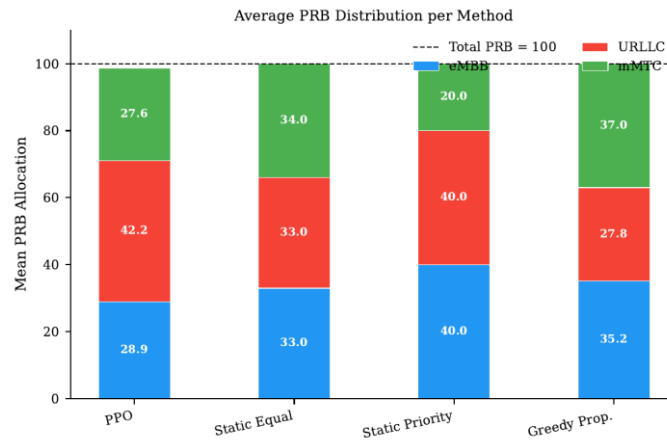


Figure 4. Mean PRB allocation per slice for each method

The PPO agent learns an asymmetric split that allocates the largest share to URLLC (≈ 42 PRBs) to protect its reliability deadline, while balancing eMBB and mMTC near the levels each needs to clear its SLA floor. This learned allocation differs markedly from the uniform split of Static Equal and the rigid reservation of Static Priority and explains how the agent simultaneously satisfies three heterogeneous contracts.

Discussion

The results support a clear conclusion: when SLA satisfaction is an explicit objective, a learned policy substantially outperforms interpretable static rules without sacrificing efficiency. The PPO agent matches the violation-free behavior of Static Equal while delivering roughly 7% more throughput, and it matches the high-throughput regime of Static Priority while reducing SLA violations by more than four orders of magnitude. The key enabler is the composite reward of Equation (8), which exposes all three objectives to the agent at once and lets it learn a load- and channel-dependent allocation rather than committing to a single fixed split. A limitation of the present study is its single-cell, single-base-station scope and its simulation-based traffic model; extending the approach to multi-cell coordination and to traces collected from real deployments is a natural direction for future work, as is the incorporation of hard constraints through constrained-RL formulations [17].

Conclusion

We presented a PPO-based, SLA-aware approach to 5G network slicing that dynamically allocates Physical Resource Blocks among eMBB, URLLC, and mMTC slices subject to per-slice contracts aligned with 3GPP TS 28.530. In a reproducible Gymnasium environment, the learned agent reduced the aggregate SLA violation rate to 2×10^{-5} while sustaining 74.7 Mbit/s of throughput and 98.6% utilization, dominating Static Equal, Static Priority, and Greedy Proportional baselines on the throughput-versus-violation frontier. The trade-off analysis confirmed that the policy reaches an operating point inaccessible to any single static rule. These findings indicate that reinforcement learning with an explicit, decomposed SLA objective is a practical path toward adaptive, contract-compliant slice management. Future work will extend the framework to multi-cell settings, real traffic traces, and constrained-RL formulations that encode SLA limits as hard constraints.

Acknowledgment

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993211 ‘Development of Solutions for 5G Networks under the ORAN Concept and Efficient Routing and Load Balancing in SDN Environments for Kazakhstan’).

References

- [1] Ericsson, "Ericsson Mobility Report," Ericsson AB, Stockholm, Sweden, Tech. Rep., Jun. 2024. [Online]. Available: <https://www.ericsson.com/en/reports-and-papers/mobility-report>
- [2] ITU-R, "IMT traffic estimates for the years 2020 to 2030," Report ITU-R M.2370-0, International Telecommunication Union, Geneva, 2015. [Online]. Available: <https://www.itu.int/pub/R-REP-M.2370>
- [3] 3GPP, "Management and orchestration; Concepts, use cases and requirements," 3GPP TS 28.530, v18.0.0, 2024. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3273>
- [4] 3GPP, "System architecture for the 5G System (5GS)," 3GPP TS 23.501, v18.0.0, 2024. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>
- [5] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, 2017, doi: 10.1109/MCOM.2017.1600951.
- [6] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 2018, doi: 10.1109/COMST.2018.2815638.
- [7] R. Li, Z. Zhao, Q. Sun, C.-L. I, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep reinforcement learning for resource management in network slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018, doi: 10.1109/ACCESS.2018.2881964.
- [8] S. Ebrahimi, F. Bouali, and O. C. L. Haas, "Resource management from single-domain 5G to end-to-end 6G network slicing: A survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 4, pp. 2836–2866, 2024, doi: 10.1109/COMST.2024.3390613.
- [9] N. Van Huynh, D. T. Hoang, D. N. Nguyen, and E. Dutkiewicz, "Optimal and fast real-time resource slicing with deep dueling neural networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1455–1470, 2019, doi: 10.1109/JSAC.2019.2904371.
- [10] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015, doi: 10.1038/nature14236.
- [11] M. R. Raza, C. Natalino, P. Ohlen, L. Wosinska, and P. Monti, "Reinforcement learning for slicing in a 5G flexible RAN," *J. Lightw. Technol.*, vol. 37, no. 20, pp. 5161–5169, 2019, doi: 10.1109/JLT.2019.2924345.
- [12] Y. Hua, R. Li, Z. Zhao, X. Chen, and H. Zhang, "GAN-powered deep distributional reinforcement learning for resource management in network slicing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 334–349, 2020, doi: 10.1109/JSAC.2019.2959185.

- [13] M. A. Ejaz, G. Wu, and T. Iqbal, "Dynamic and efficient resource allocation for 5G end-to-end network slicing: A multi-agent deep reinforcement learning approach," *Int. J. Commun. Syst.*, vol. 37, no. 17, art. e5916, 2024, doi: 10.1002/dac.5916.
- [14] N. Saha, M. Zangoeei, M. Golkarifard, and R. Boutaba, "Deep reinforcement learning approaches to network slice scaling and placement: A survey," *IEEE Commun. Mag.*, vol. 61, no. 2, pp. 82–87, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10004899>
- [15] M. Sulaiman, A. Moayyedi, M. Ahmadi, M. A. Salahuddin, R. Boutaba, and A. Saleh, "Coordinated slicing and admission control using multi-agent deep reinforcement learning," *IEEE Trans. Netw. Serv. Manage.*, vol. 20, no. 2, pp. 1110–1124, 2023, doi: 10.1109/TNSM.2022.3222589.
- [16] M. Khani, S. Jamali, M. K. Sohrabi, M. M. Sadr, and A. Ghaffari, "Slice admission control in 5G cloud radio access network using deep reinforcement learning: A survey," *Int. J. Commun. Syst.*, vol. 37, no. 13, art. e5857, 2024, doi: 10.1002/dac.5857.
- [17] M. Sulaiman, M. Ahmadi, M. A. Salahuddin, R. Boutaba, and A. Saleh, "Generalizable resource scaling of 5G slices using constrained reinforcement learning," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, Miami, FL, USA, 2023, pp. 1–9, doi: 10.1109/NOMS56928.2023.10154267.
- [18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017, doi: 10.48550/arXiv.1707.06347.
- [19] R. Barker, A. Ebrahimi Dorcheh, T. Seyfi, and F. Afghah, "REAL: Reinforcement learning-enabled xApps for experimental closed-loop optimization in O-RAN with OSC RIC and srsRAN," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2025, doi: 10.1109/ICCWorkshops67674.2025.11162144.
- [20] N. Shezi, B. Nleya, and B. Pule, "Slice-aware and computationally efficient resource orchestration for converged mmWave–PON O-RAN: A reward-shaped PPO approach for joint DBA and PRB allocation," *Telecom*, vol. 7, no. 3, art. 75, 2026, doi: 10.3390/telecom7030075.
- [21] Y. Azimi, S. Yousefi, H. Kalbkhani, and T. Kunz, "Energy-efficient deep reinforcement learning assisted resource allocation for 5G-RAN slicing," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 856–871, 2022, doi: 10.1109/TVT.2021.3128513.
- [22] S. Choi, S. Choi, G. Lee, S.-G. Yoon, and S. Bahk, "Deep reinforcement learning for scalable dynamic bandwidth allocation in RAN slicing with highly mobile users," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 576–590, 2024, doi: 10.1109/TVT.2023.3302416.
- [23] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: <https://jmlr.org/papers/v22/20-1364.html>
- [24] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 3GPP TS 36.213, v17.0.0, 2022. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2427>