Learning-Aided Multi-RAT Operation for Battery Lifetime Extension in LPWAN Systems

Martin Stusek[1,2], Dmitri Moltchanov[2], Pavel Masek[1], Jiri Hosek[1], Sergey Andreev[2], and Yevgeni Koucheryavy[2]

[1]Department of Telecommunications, Brno University of Technology, Brno, Czech Republic
[2]Unit of Electrical Engineering, Tampere University, Finland
✉Contact author's e-mail: masekpavel@vutbr.cz

***Abstract*—End-device (ED) lifetime is considered to be a crucial design factor in radio systems for massive machine-type communications. This parameter is heavily impacted by the continuously changing propagation conditions between the ED and the base station. In this paper, to extend the ED lifetime, we consider equipping a single ED with multiple low-power wide-area network (LPWAN) technologies to dynamically select the one with lower energy consumption. To facilitate this process, we propose to employ reinforcement learning (RL) algorithms. Assessing the resultant performance, we conduct two large-scale measurement campaigns that characterize the ED power consumption and the time-dependent propagation conditions for NB-IoT, Sigfox, and LoRaWAN technologies. Our numerical results demonstrate that the designed schemes effectively reduce ED power consumption by timely reacting to the varying radio conditions. Consequently, the ED lifetime expectancy is prolonged by around 10%. For instance, the Thompson sampling technique delivers the most consistent results by outperforming its counterparts and allowing to exploit up to 99% of the theoretical gains while converging over only 25-50 samples.**

***Index Terms*—End-device lifetime; Energy consumption optimization; LPWAN; Multi-armed bandit; Multi-RAT; Reinforcement learning**

I. Introduction

Over the recent decade, the vision of billions of interconnected machines across the landscape of the Internet of Things (IoT) has initiated the research and development work toward heterogeneous wireless communication systems [1]. Since the general principles of how to assign the system resources and transmit the data have remained nearly unchanged in the legacy 3GPP-defined systems, i.e., 2G, 3G, 4G, the next-generation mobile systems 5G and beyond 5G (B5G) are expected to fulfill the highly challenging needs of enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC) [2], [3]. Being one of the key pillars of the (B)5G, the mMTC defines novel requirements that are difficult to satisfy by only relying on the legacy wireless systems.

In contrast to the legacy solutions, a new class of Radio Access Technologies (RATs), i.e., low-power wide-area networks (LPWANs) is gaining momentum due to its communication parameters designed from scratch to satisfy the diverse communication requirements, e.g., the message delivery over less than 10 s, with message loss probability as low as 10% in the case of remote metering [4]. Currently, the scope of today's IoT-ready devices mostly pertains to the utilization of a single RAT. On the one hand, this helps reduce the price point of the devices, but on the other hand it limits the possible communication scenarios – as it relies on the selected LPWAN technology [5].

The concept of combining several LPWAN technologies within one multifunctional end-device (ED) opens the door to addressing the needs of next-generation IoT devices, which will operate in the emerging 5G and B5G networks [6]. The hypothesis behind the proposed study is that propagation conditions between the ED and the LPWAN base station (BS) may change dramatically over the ED lifetime [7]. Various environmental changes, including natural and human-made, such as weather conditions, construction works, and large infrastructural changes, may differently affect the operating bands of the technologies at the timescales from seconds to years. These phenomena are difficult to predict at the time of ED deployment, and their ultimate effect is not straightforward to assess in the ultra-high frequency (UHF) band utilized by LPWAN technologies. Although the use of multi-RAT sensors adds to the complexity and the cost of EDs, it may have a positive effect on the sensor lifetime [8].

In this paper, we envision to utilize multiple RATs at a single ED and dynamically switch between them by employing the best technology available currently in terms of the propagation conditions. Since the propagation conditions are not known in advance, the selection of the best RAT has to be conducted automatically in response to environmental changes. One of the potential tools that allows for a dynamic adaptation is reinforcement learning (RL), where the system continuously assesses the environment and assigns the weights to different options by attempting to maximize its reward. In our case, the reward is related to the minimization of the ED power consumption.

To quantify the trade-offs between the performance metrics of interest instead of relying solely on simulation frameworks, we perform two large-scale measurement campaigns that characterize the ED power consumption and time-dependent reference signal received power (RSRP)/received signal strength indicator (RSSI). We focus on three widely utilized LPWAN technologies, which are Narrowband IoT (NB-IoT), Sigfox, and LoRaWAN. While LPWAN technologies from IEEE family are kept aside due to lack of market-available

hardware [9]. We then proceed with modeling time-dependent channel behavior with a Markov chain framework. Finally, we demonstrate the utilization of RL-based techniques for battery lifetime extension where EDs are equipped with multiple LP-WAN interfaces by considering and comparing the following approaches: ε-greedy, upper confidence bound (UCB), and Thompson sampling.

Our main contributions are:

- formalization of RL-based approaches for battery lifetime extension of LPWAN EDs supporting multi-RAT capabilities;
- relying on RL-based operation to extend the battery lifetime by 10% when all three considered LPWAN technologies are utilized simultaneously;
- showing that Thompson sampling outperforms ε-greedy and UCB options in the considered scenarios by exploiting up to 99% of the theoretical gains and reaching this regime in only 25-50 samples.

The rest of this text is organized as follows. In Section II, we introduce our system model with respect to the multi-RAT functionality. Further, in Section III, we characterize power consumption of the considered LPWAN technologies. We report time-dependent channel characteristics and the associated models in Section IV. Implementation of the RL algorithms is sketched in Section IV together with illustrative numerical results. Finally, conclusions are drawn in the last section.

II. Environment and Assumptions

This section discusses the LPWAN technologies selected for our measurement campaign. Further, the multi-RAT functionality is described with respect to the metrics of interest.

A. Considered mMTC Technologies

In this work, we consider three LPWAN technologies, which are globally-known LPWAN representatives, i.e., NB-IoT, Sigfox, and LoRaWAN. As it is detailed in Table I, even though these LPWAN technologies have similar targets, their communication capabilities are different.

The proprietary solutions Sigfox and LoRaWAN operate in the license-exempt industrial, scientific, and medical (ISM) band by relying on ALOHA-type channel access [10]. Since both technologies use the license-free spectrum, anyone can construct a low-budget ED and deploy their solution. A possible drawback of the ISM band utilization is the need to abide by the duty-cycle restrictions defined in the frequency regulations of the selected countries.

Contrary to the aforementioned technologies, NB-IoT utilizes the licensed frequency bands, where the available network resources can be assigned to the EDs. This approach permits to control collisions as well as perform uplink and downlink data transmissions efficiently. The downsides of this 3GPP-ratified technology relate to the need for network synchronization, which increases the power consumption and the operating times [11].

TABLE I: Key parameters of LPWAN technologies.

	LoRaWAN	**Sigfox**	**NB-IoT**
Coverage (MCL)	157 dB	162 dB	164 dB
Technology	Proprietary	Proprietary	Open LTE
Spectrum	Unlicensed	Unlicensed	Licensed
Frequency	433, 868, 915 MHz	868, 915 MHz	700–2100 MHz
Bandwidth	125, 250, 500 kHz	100, 600 Hz	200 kHz
Max. ERP	14 dBm	14 dBm	23 dBm
Downlink data rate	0.3-50 kbps	0.6 kbps	$<$ 300 kbps
Uplink data rate	0.3-50 kbps	0.1-0.6 kbps	0.3-62.5 kbps[1]
Max. message size UL	242 B	12 B	1600 B
Max. message size DL	242 B	8 B	1600 B
Battery life	10+ years	10+ years	10+ years
Module cost	6 $	3 $	12 $
Security	AES-128	AES-128	LTE Security

B. Deployment and Multi-RAT Functionality

From the discussion above, we conclude that the distinct features of the considered LPWAN technologies make them attractive for specific applications instead of generic IoT scenarios. This supports our initial premise to enable a multi-RAT setup that employs several complementary technologies. Not limited to more diversity in the utilized frequency spectrum, the selected technologies also differ in the respective business models, thus underpinning further possible benefits from combining them.

Our measurement campaign has been carried out in the city of Brno, Czech Republic. Specifically, two different locations were selected: (i) Brno University of Technology (BUT) campus on the outskirts of the city, and (ii) an apartment located in the city center. The primary metric of interest targeted by our study is the ED battery lifetime. We assess it by utilizing an intermediate parameter – ED power consumption required for communication in a multi-RAT environment.

The implementation of multi-RAT functionality directly affects the communicating device as well as impacts the power consumption of the ED. The possible operation modes in the case of multi-RAT devices include:

- *Parallel*: the device is capable of transmitting data over various RATs at the same time. The main drawback is in increased power consumption for all operating states. In addition, the processing ED resources to enable multiple transmission pipelines increase the device cost.
- *Selective*: even though the device is equipped with several radio interfaces, it utilizes only one of them at a time. This approach allows for a dynamic adjustment of the communication technology based on the radio conditions. Compared to the "parallel" setup, it is also expected to be less demanding with respect to power and processing capabilities.
- *Sequential*: a combination of the above options, which supports the predefined parallel operations (network synchronization), but selects one technology as the communication enabler, i.e., sends data over one technology at a time.

C. Approach at a Glance

In this study, we consider the "selective" multi-RAT functionality. To assess the ED lifetime gains enabled by this multi-RAT operation, we utilize the following methodology: (i) we characterize the energy efficiency of the considered LPWAN technologies in different propagation conditions, (ii) to represent realistic propagation conditions of each LPWAN technology, we then conduct the RSRP/RSSI measurement campaign, (iii) the resulting data is further used to identify the associated time-dependent propagation models using a doubly-stochastic Markov chain framework, (iv) by employing the developed channel models and energy consumption assessments, we then formulate and apply the multi-armed bandit (MAB) RL framework with several enhancements.

III. Power Consumption Characterization

To predict the expected battery life of multi-RAT LPWAN devices, it is essential to measure the energy consumption of each technology in all the operating states. To this aim, we conducted extensive energy consumption measurements covering all the considered LPWAN technologies in the laboratory environment. This setup allowed us to measure the consumption of modules at signal levels from -68 to -133 dBm (it covers all ECL classes, i.e., ECL 0 – ECL 2).

The desired signal levels were achieved via a step attenuator positioned between the remote radio unit (RRU) and the measured NB-IoT module placed in the radio frequency (RF) shield box, see Fig. 1. The consumption was measured via a power analyzer Agilent N6705A, which allows for the sampling period of as low as 0.0248 ms. Due to the limitations of Sigfox technology, we measured the consumption for 12 B message transmission. The measurements at each signal level (12 signal levels in total) were repeated 10 times for each technology to gain sufficient confidence bounds.

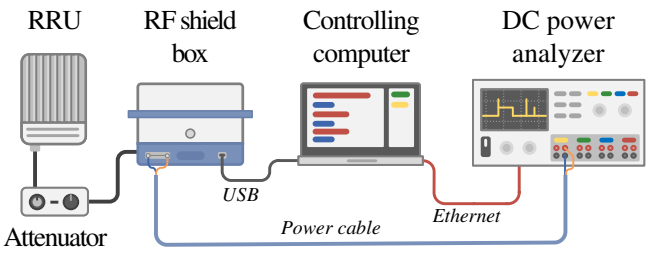


Fig. 1: Workplace for NB-IoT energy measurements.

In the case of Sigfox and LoRaWAN, the measurements were simplified. Since these technologies do not require network registration prior to a message transmission, energy consumption is not dependent on the signal level. Hence, we moved the modules out of the shield box and conducted the energy measurements. As a representative of NB-IoT technology, we selected a communication module SARA N210 by uBlox company that implements the Rel. 13 version of the standard. For LoRAWAN, the module designated as Microchip RN2483 was used. Finally, for the Sigfox technology, we utilized the S2-LP communication module produced by STMicroelectronics company.

A. Measurement Results

Power consumption during a message transmission depicted in Fig. 2 clearly illustrates that for NB-IoT in the class ECL 0, the consumption remains almost constant. A first visible increase is noticeable in ECL 1, followed by a steep rise in ECL 2. A side-by-side comparison of the ECL classes reveals that the consumption in ECL 2 can be 15 times higher as compared to ECL 0. The 5th and 95th percentiles represented by the whiskers also indicate a higher dispersion of the results. This difference is primarily caused by message retransmissions, which play a crucial role under poor signal conditions and may vary significantly. In the case of LoRaWAN, the consumption utilizing SF 12 is comparable to that of NB-IoT operating in ECL 0. On the other hand, Sigfox results resemble those for the NB-IoT energy consumption in ECL 1.

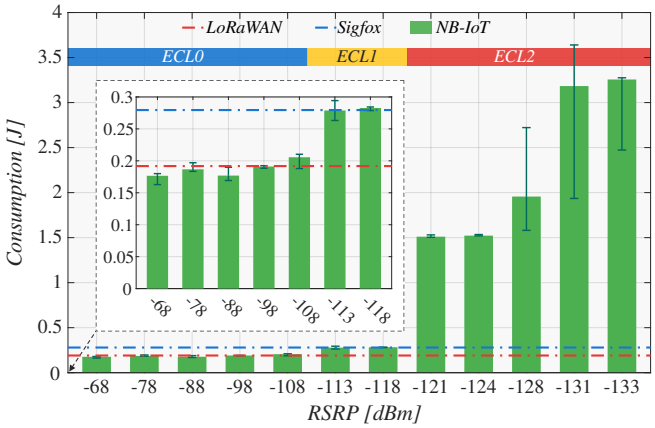


Fig. 2: Results of power consumption measurements.

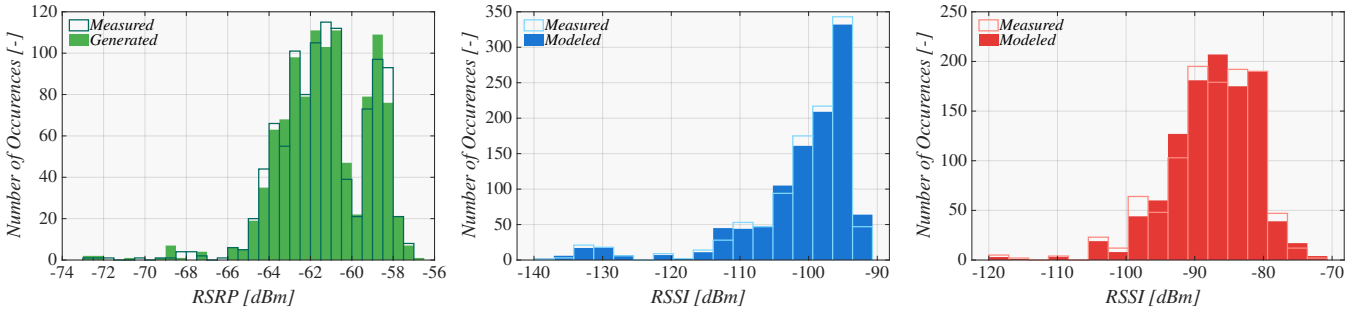
IV. Time-Dependent Propagation Modeling

In this section, we first describe our measurement campaign and then proceed by specifying the time-dependent propagation models based on our doubly-stochastic Markov chain framework.

A. Equipment and Measurement Methodology

For the purposes of time-dependent signal characteristics, we designed two multi-RAT evaluation boards. The first one was located in the office at BUT premises, whereas the second prototype was positioned in a private apartment close to the Brno city center. Both devices were set to convey 12 B messages with one hour period. The same message was always sent via all the radio interfaces, and the available statistics were collected. The devices operated in a sequential mode to reduce the possibility of interference, since the technologies use the neighboring frequency bands (868 MHz for Sigfox and LoRaWAN, 800 MHz for NB-IoT).

Throughout the two-month measurement campaign, we acquired more than 1400 messages for each LPWAN technology from every sensor. All radio solutions utilized the publicly



(a) NB-IoT (b) Sigfox (c) LoRaWAN

Fig. 3: Relative frequency histogram of RSRP/RSSI samples.

available networks with a multi-gateway setup. Hence, the measurements reflected real-world conditions where a sensor is employed for the purposes of production deployment. Concerning the signal quality parameter, we utilized RSSI for Sigfox and LoRaWAN technologies. For NB-IoT, we focused on RSRP as it potentially excludes noise and interference from other sectors.

B. Time-Dependent Propagation Model

We start by identifying the classes of models that can be utilized for RSRP/RSSI assessment. To this aim, we observe the first- and second-order statistical characteristics illustrated in Fig. 3 and Fig. 4, i.e., the histogram and autocorrelation function (ACF). All data presented in this section was acquired from the sensor deployed at BUT premises. As one may observe, the histogram has a specific structure that cannot be modeled accurately by any known distribution.

In its turn, the ACF is characterized with a near-exponential decay by eventually approaching zero. These observations allow us to make the following conclusions: (i) the considered RSRP/RSSI process is ergodic in nature, (ii) doubly-stochastic Markov chain framework (also known as hidden Markov model) is one of the candidates for the target assessment [12]–[14]. This framework offers a balance between the modeling accuracy and the simplicity of fitting algorithms.

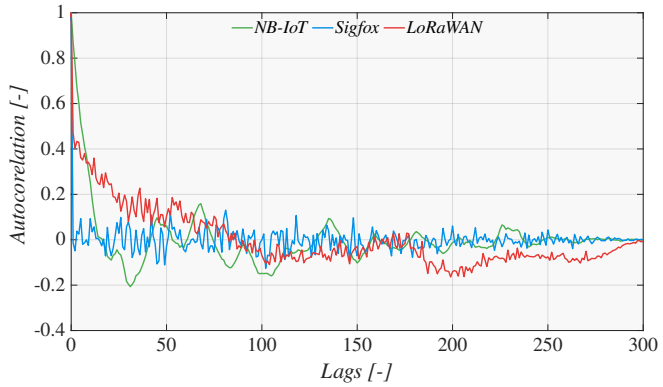


Fig. 4: ACFs for considered measurements.

To parametrize the doubly-stochastic Markov model, one needs to determine the number of states, N, and then estimate the transition probabilities from the current i to the next j state

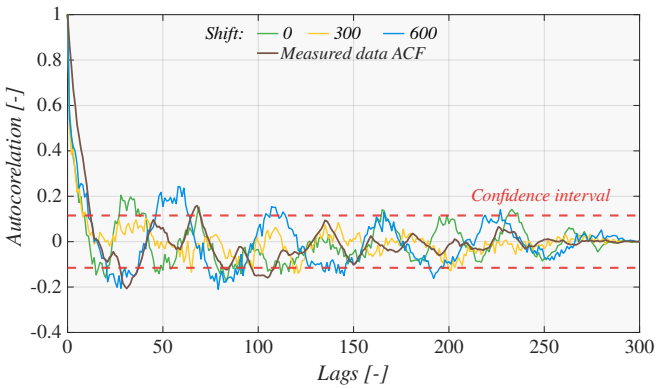


Fig. 5: Comparison of ACFs for NB-IoT.

p_{ij}, $i, j = 1, 2, \ldots, N$ as well as the conditional probability mass functions (PMFs) associated with each state, $f_i(j)$, $i = 1, 2, \ldots, N$, $j \geq 0$. There is a number of algorithms for fitting the doubly-stochastic Markov models for statistical data. We utilize the following procedure.

To detect the number of states, we apply the kernel density estimation (KDE) to cluster the data [15]. This process consists of two steps: (i) estimation of probability density function (PDF), and (ii) data clusterization based on local maxima. The PDF of the samples is calculated as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{1}$$

where n is the sample size, h is the bandwidth, x is the actual value, and x_i represents the input samples. As the kernel density estimator K, we used the normal kernel function. The states of the Markov chain are derived from the local maxima of the resulting KDE. Each local maximum represents one boundary of the Markov chain state.

Once the number of states is determined, we proceed to derive the transition probabilities p_{ij}, $i, j = 1, 2, \ldots, N$ and the PMFs associated with each state, $f_i(j)$, $i = 1, 2, \ldots, N$, $j \geq 0$ by using conventional statistical methodologies. We thus define the state boundaries between the states and then calculate the number of state changes for the particular values of i and j between the previous and the current value in the

trace. Finally, we divide number of state changes by the trace size.

Aiming to assess the performance of the proposed model, we plot the histograms and ACFs of the statistical and model data, see Figs. 3a and 5. The histograms represent the results for all the LPWAN technologies, but the ACF only shows the NB-IoT data since other options behave similarly. As one may observe, the histograms match tightly. Particularly, the χ^2 goodness-of-fit test performed with the level of significance set to 0.05 confirms that both samples belong to the same distribution. When analyzing the results, it is evident that the exponentially decaying behavior of the ACF is also well captured. This allows to conclude that the proposed approach can be utilized for time-dependent RSRP/RSSI modeling. Similar conclusions are made for other technologies.

V. Implementation and Numerical Results

In this section, we first offer the implementation details for the MAB-based RL techniques and then continue to report our selected numerical results.

A. Implementation of RL Policies

In our work, we employ four different types of policies for MAB problem-solving. Namely, our selected policies include ε-greedy, weighted average, UCB, and Thompson sampling.

1) ε-greedy: is the simplest method of assessing the exploration-exploitation problem. The algorithm explores, which arm to pull randomly with the probability given by the parameter ε. In the remaining time $(1-\varepsilon)$, the algorithm attempts to gain the highest reward from each action, i.e., operates in the exploitation mode. To be able to pull the best arm, the algorithm keeps the averages of all the observed rewards from each arm. These averages are calculated in an incremental manner to save the computational resources. The value of $k+1$ action Q_{k+1} can be produced as:

$$Q_{k+1} = Q_k + \frac{1}{k+1}\left[r_{k+1} - Q_k\right], \tag{2}$$

where k is the order of the actual step, r_{k+1} is the actual reward, and Q_k is the average of the first k rewards [16].

2) Weighted average: represents a modification of the ε-greedy algorithm suitable for a non-stationary environment. In this case, the ε-greedy algorithm performs poorly since it computes the mean based on all the previous samples. In a non-stationary environment, it is reasonable to weigh the recent rewards more heavily than the old ones. With this modification, the averaged mean is calculated as:

$$Q_{k+1} = Q_k + \alpha\left[r_{k+1} - Q_k\right], \tag{3}$$

where α $(0 < \alpha \leq 1)$ represents the step-size constant [16].

3) UCB: While the ε-greedy algorithm indiscriminately selects an action without a preference among the arms, it may leads to inefficient exploration. The UCB algorithm, on the other hand, favors the exploration of actions with high potential to reach the optimal value. The actions are selected based on the proximity to the maximum and the uncertainty of a selection as

$$A_t = argmax_a\left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}}\right], \tag{4}$$

where $c > 0$ controls the degree of exploration and $N_t(a)$ denotes how many times action a has been triggered prior to time t [17].

4) Thompson sampling: is similar to UCB in that it utilizes an approach to arm selection for increased exploration efficiency. Accordingly, Thompson sampling draws from the posterior predictive Beta distribution of each choice by using a random uniform variable. This enables a non-optimal distribution to be sampled with varying frequency. Since the distribution certainty increases, the probability of the choice being made decreases dynamically while balancing the need for more information by making the currently optimal choice [18].

B. Average Rewards

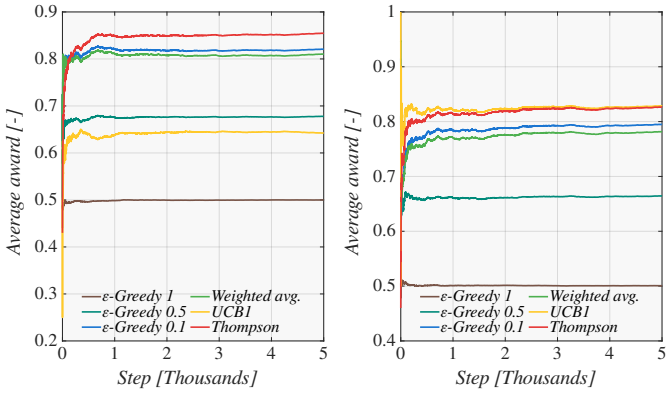
As the first step in evaluating the policies, we conducted 200 RL runs with the algorithms on 5000 samples. The samples were derived by using the developed Markov chain models. These samples were further employed to construct a rewards matrix utilized as an input to the simulation runs. The best performing LPWAN technology in each step was rewarded with value 1, second with 0.5, and, finally, the technology with the highest energy consumption for a message transmission received 0. The RL process was conducted based on data from both sensors, i.e., the one at BUT campus and another one in the Brno city center (BCC).

The results for both data sets are depicted in Fig. 6. For the BUT data, Thompson sampling provides the best results. Further, the performance of the UCB algorithm does not meet the expectations with only 0.63 average rewards, whereas the ε-greedy and weighted average both achieve the values of over 0.8. In this case, UCB most probably identified a local optimum solution. As expected, the ε-greedy 1 (always explore) scheme settled for 0.5 average rewards, which reflects a random selection.

However, for the BCC sensor the situation is different, and the UCB provides the best results. Thompson sampling is then in the second place, which makes this approach a reliable option. Order of the remaining techniques repeats the case of the BUT sensor data, with slightly decreased average rewards. It is also important to note that the RL-based approach allows to exploit up to 85% of the theoretical gains. Furthermore, each algorithm is able to achieve 90% of its maximum average rewards during the initialization phase in less than 50 messages. With the UCB algorithm, it is possible to reach this value with only 25 messages. Such a short convergence time is especially beneficial for battery-powered devices.

C. Battery Life Expectancy

In the last step of our RL-aided evaluation, we extend the length of the samples by making the cumulative consumption

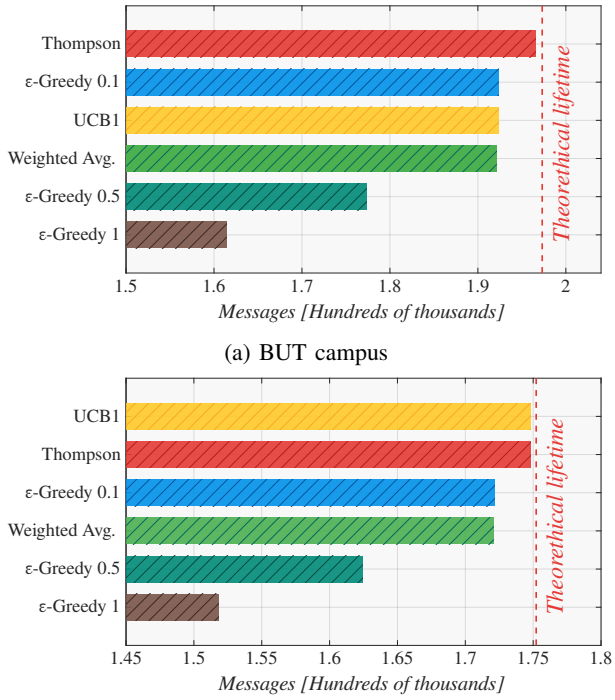


(a) BUT campus (b) Brno city center

Fig. 6: Average rewards of reinforcement learning policies.

of the transmitted messages equal to the nominal capacity of the device battery. For this purpose, we select a commonly used lithium battery Saft LS 14500 with the capacity of 2.6 Ah (33696 J).

The BUT results demonstrate that Thompson sampling provides the longest battery life, see Fig. 7. If one compares this result with the theoretical maximum where is the knowledge of the best performing technology at each step, the performance of Thompson sampling is only 0.5% lower. UCB and ε-greedy with $\varepsilon = 0.1$ and simple weighted average algorithms display a similar performance by only losing 5-10% to the Thompson sampling technique.



(a) BUT campus

(b) Bno city center

Fig. 7: Battery life expectancy for all studied policies.

Recalling the average rewards, for the BCC sensor data, the first place is kept by the UCB algorithm. It outperforms Thompson sampling by about 500 messages and comes close to the theoretical upper bound in terms of the sensor lifetime. Further, the Thompson sampling technique is on a par with UCB by showing extremely promising results, whereas the ε-greedy algorithm, with both of the considered ε values, provides significantly inferior results.

Finally, we note that for the ε-greedy algorithm, the choice of ε value is crucial. As illustrated in Fig. 7, for both of the considered sensor data alternatives selecting ε close to unity leads to extreme performance degradation as the algorithm is then forced to explore new RAT options. In this case, the utilization of non-parametric Thompson sampling and UCB techniques may allow for a preffered solution.

VI. Conclusions

In this paper, by aiming to improve the ED lifetime in LPWAN systems under varying propagation conditions, we considered the utilization of multi-RAT options at a single ED and the dynamic switching between them throughout its operation. To facilitate the radio selection process, we proposed to employ MAB-based RL techniques. Accordingly, several schemes have been implemented and assessed by using two-month practical measurement campaigns centered at ED power consumption and time-dependent channel characterization.

Our obtained numerical results indicated that the considered RL-based techniques may reduce the impact of multi-RAT operation on the power consumption since the proposed approach reacts based on the dynamic radio conditions and helps extend the ED battery lifetime expectancy by around 10%. It is important to note that each studied algorithm is able to achieve 90% of its maximum average rewards during the initialization phase in less than 50 messages. For both EDs in question, it was confirmed that Thompson sampling offers the most consistent results. Even in its worst case, the cumulative increase in the power consumption does not exceed 0.5%.

Acknowledgment

For the research, the infrastructure of the SIX Center was used. The described research was financed by the Technology Agency of the Czech Republic project No. TN01000007. TAU work was supported by the Business Finland Project 5G-FORCE.

References

[1] K. Samdanis and T. Taleb, "The Road beyond 5G: A Vision and Insight of the Key Technologies," *IEEE Network*, vol. 34, no. 2, pp. 135–141, 2020.

[2] D. Solomitckii, A. Orsino, S. Andreev, Y. Koucheryavy, and M. Valkama, "Characterization of mmWave Channel Properties at 28 and 60 GHz in Factory Automation Deployments," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.

[3] I. B. F. de Almeida, L. L. Mendes, J. J. Rodrigues, and M. A. da Cruz, "5G Waveforms for IoT Applications," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2554–2567, 2019.

[4] J. Finnegan and S. Brown, "A comparative survey of LPWA networking," *arXiv preprint arXiv:1802.04222*, 2018.

[5] R. Mozny, M. Stusek, P. Masek, K. Mikhaylov, and J. Hosek, "Unifying Multi-Radio Communication Technologies to Enable mMTC Applications in B5G Networks," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*. IEEE, 2020, pp. 1–5.

[6] H. Tabassum, M. Salehi, and E. Hossain, "Fundamentals of Mobility-Aware Performance Characterization of Cellular Networks: A Tutorial," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2288–2308, 2019.

[7] S. Kavuri, D. Moltchanov, A. Ometov, S. Andreev, and Y. Koucheryavy, "Performance Analysis of Onshore NB-IoT for Container Tracking During Near-the-Shore Vessel Navigation," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2928–2943, 2020.

[8] K. Mikhaylov, V. Petrov, R. Gupta, M. A. Lema, O. Galinina, S. Andreev, Y. Koucheryavy, M. Valkama, A. Pouttu, and M. Dohler, "Energy Efficiency of Multi-Radio Massive Machine-Type Communication (MR-MMTC): Applications, Challenges, and Solutions," *IEEE Communications Magazine*, vol. 57, no. 6, pp. 100–106, 2019.

[9] A. Ometov, N. Daneshfar, A. Hazmi, S. Andreev, L. F. D. Carpio, P. Amin, J. Torsner, Y. Koucheryavy, and M. Valkama, "System-level Analysis of IEEE 802.11ah Technology for Unsaturated MTC Traffic," *International Journal of Sensor Networks*, vol. 26, no. 4, pp. 269–282, 2018.

[10] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low Power Wide Area Networks: An Overview," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 855–873, 2017.

[11] W. Yang, M. Wang, J. Zhang, J. Zou, M. Hua, T. Xia, and X. You, "Narrowband Wireless Access for Low-Power Massive Internet of Things: A Bandwidth Perspective," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 138–145, 2017.

[12] T. Rolski, "Upper Bounds for Single Server Queues With Doubly Stochastic Poisson Arrivals," *Mathematics of Operations Research*, vol. 11, no. 3, pp. 442–450, 1986.

[13] S. B. Slimane and T. Le-Ngoc, "A Doubly Stochastic Poisson Model for Self-similar Traffic," in *Proceedings IEEE International Conference on Communications ICC'95*, vol. 1. IEEE, 1995, pp. 456–460.

[14] D. Moltchanov, Y. Koucheryavy, and J. Harju, "The Model of Single Smoothed MPEG Traffic Source Based on the D-BMAP Arrival Process with Limited State space," in *Proc. of ICACT*, 2003, pp. 55–60.

[15] Y. Huang, X. Chen, and W. B. Wu, "Recursive Nonparametric Estimation for Time Series," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1301–1312, Feb 2014.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[17] T. Lattimore and C. Szepesvári, "Bandit algorithms," *Cambridge University Press (preprint)*, 2018.

[18] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "A Tutorial on Thompson Sampling," *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018. [Online]. Available: http://dx.doi.org/10.1561/2200000070