Reinforcement Learning Empowered Massive IoT Access in LEO-based Non-Terrestrial Networks

Ju-Hyung Lee[†‡], Dheeraj Panneer Selvam[†], Andreas F. Molisch[†], and Joongheon Kim[‡]
[†]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA
[‡] School of Electrical and Computer Engineering, Korea University, Seoul, Korea
juhyung.lee@usc.edu, dpanneer@usc.edu, molisch@usc.edu, joongheon@korea.ac.kr

***Abstract*—Low-Earth orbit (LEO) satellite (SAT) networks exhibit ultra-wide coverage under time-varying SAT network topology. Such wide coverage makes the LEO SAT network support the massive IoT, however, such massive access put existing multiple access protocols ill-suited. To overcome this issue, in this paper, we propose a novel contention-based random access solution for massive IoT in LEO SAT networks. Not only showing the performance of our proposed approach (see, Table II), but we also discuss the issue of scalability of deep reinforcement learning (DRL) by showing the convergence behavior (see, Table III and IV).**

***Index Terms*—LEO satellite network, massive IoT, random access, reinforcement learning, 6G.**

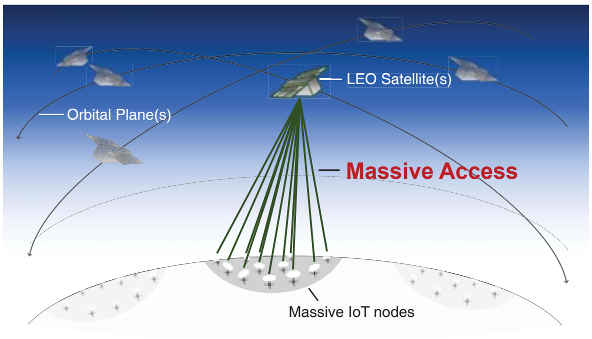

Figure 1: An illustration of massive IoT access in LEO satellite networks.

I. Introduction

Low-Earth Orbit (LEO) satellite (SAT) constellation has emerged as a dominant paradigm for sixth generation (6G) communications. Several industry projects, including SpaceX, OneWeb, and Amazon, spur on this trend by launching thousands of LEO SATs and their realistic blueprints [1], [2]. In particular, more than thousands of deployed SATs by *Starlink* [3] have currently supported more than 30 countries in North America, Europe, and Oceania continents [4], [5]. Such a new type of wireless connectivity, a mega-constellation of LEO SATs, has a great potential in provisioning fast and reliable connectivity to ground users anywhere in the globe, including ocean, rural areas, and disaster sites.

Thanks to the ultra-wide coverage and periodical orbital move of the LEO SAT network, LEO SAT-based NTN can support numerous scattered internet of things (IoT) on Earth, albeit such massive access by scattered innumerable ground nodes disallows sophisticated central coordination among SAT BSs and users in real-time under limited communication and computing resources. Furthermore, as opposed to fixed terrestrial BSs, LEO SAT is a mobile base station, requiring location-specific resource management. Accordingly, existing model-based and standardized access protocols, e.g., slotted ALOHA and random access channel (RACH), cannot flexibly optimize their operations without incurring severe protocol fragmentation, not to mention a significant effort in protocol standardization for a myriad of possible scenarios. To overcome the fundamental challenges in the massive IoT-NTN scenario [6], in this article, we propose a novel contention-based access (RA) protocol.

II. System Model

A. Geometry of LEO SAT

We consider sets $\mathcal{I}$ of LEO SATs and a set $\mathcal{K}$ of IoT nodes deployed on the ground inside an area A. The position of IoT node $k \in \mathcal{K}$ is expressed as a 3-dimensional real vector on Cartesian coordinates denoted by $\boldsymbol{q}_k = (q_k^x, q_k^y, q_k^z) \in \mathbb{R}^3$, and similarly, the position and velocity of SAT $i \in \mathcal{I}$ at time $t \geq 0$ is denoted by $\boldsymbol{q}_i(t) = (q_i^x(t), q_i^y(t), q_i^z(t)) \in \mathbb{R}^3$ and $\boldsymbol{v}_i(t) = (v_i^x(t), v_i^y(t), v_i^z(t)) \in \mathbb{R}^3$, respectively, for all $i \in \mathcal{I}$. Suppose the number of IoT nodes is given as $|\mathcal{K}| = K$, and assume all SATs are moving in uniform circular motion with the same orbital period T, while the arc length between any two neighboring SATs on the same orbital plane is equal to each other. Also, consider that time is discretized in slots of length τ and let $\boldsymbol{q}_i[0]$ be the initial position of the SAT $i \in \mathcal{I}$ at time $t = 0$. Then, by following the discrete-time state-space model [7], [8], the position of SAT i at time $t = m\tau$ can be expressed as

$$\boldsymbol{q}_i(m\tau) \approx \boldsymbol{q}_i(0) + \tau \sum_{m'=1}^{m} \boldsymbol{v}_i(m'\tau). \tag{1}$$

B. Access Scenario

Consider an networks scenario of NB-IoT NTN (IoT-NTN), where massive IoT nodes attempt to access to LEO SAT networks for radio resource grant. For the sake of convenience, we consider that IoT always have intentions to access at every opportunity they have; and we suppose that each UT has information of the periodic position of SATs on each orbital plane and attempts to access only to the closest SAT on each orbital plane.

At each access opportunity, each IoT nodes chooses whether to access or backoff. Such set of actions is simply denoted by $\{0, 1\}$. The RA action of IoT node $k \in \mathcal{K}$ at access opportunity n is denoted by

$$w_k[n] \in \{0, 1\}. \tag{2}$$

Note that $w_k[n] = 0$ means that the IoT node j does not access at the n-th opportunity and waits for the next one, that is *backoff*. Besides, those who attempt to access choose a preamble uniform-randomly over the predefined signature sequence index set $s_k[n] = \{1, 2, \ldots, R\}$. Note that each preamble is associated with R resources that the SATs can grant during the data transmission duration; contention may occur when multiple terminals select the same PRACH.

Once the IoT node decides the access action, $w_k[n]$, the node randomly selects the preamble signatures and then transmits the PRACH preamble sequence, as in 5G NR and 4G LTE/LTE-A [9], [10]. Recall that when a node transmit a PRACH Preamble, it transmits with a specific pattern and this specific pattern is called a *signature*. In each LTE cell, total 64 preamble signatures are available and UE select randomly one of these signatures [11].

C. Figure of Merit

To evaluate the performance of the proposed scheme, the successful access rate S is our main figure of merit.

The collision information, $c_k[n]$, represents if a collision occurs in access action of IoT j for time slot n, which is given by

$$c_k[n] = \begin{cases} 1, & w_k[n] \neq a_{k'}[n], s_k[n] \neq s_{k'}[n], \forall k^{'} \in \mathcal{K}/j, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The collision rate is defined as

$$C = \sum_{n=1}^{N} \sum_{k \in \mathcal{K}}^{N} c_k[n]. \tag{4}$$

We define the access indicator of IoT $j \in \mathcal{K}$'s random access at random access opportunity $n \in \{1, \ldots, N\}$ as

$$\eta_k[n] = \begin{cases} (1 - c_k[n]), & w_k[n] \neq 0, \\ 0, & w_k[n] = 0, \end{cases} \tag{5}$$

. Let S be the number of successful accesses of all IoTs out of N access attempts, which is given as

$$S = \frac{1}{|\mathcal{K}|} \sum_{n=1}^{N} \sum_{k \in \mathcal{K}} \eta_k[n]. \tag{6}$$

III. Reinforcement Learning for Massive IoT Access in LEO Satellite Networks

The IoT nodes on the ground attempts to access an SAT by the contention-based RA and then transmits a data frame only when it successfully accesses to intended LEO SAT. Here, we aims to minimize the collision under the constraints related to the practical conditions of LEO SAT networks. Recently, a new model-free protocol has been investigated, using model-free DRL algorithms, which can be an alternative solution for such a time-varying network topology [12], [13]. To optimize the access protocol for massive access scenario in IoT-NTN, we used a centralized type DRL method. We discuss a Markov decision process (MDP) model, that reflects our network scenario, in the following.

Table I: Simulation parameters.

Parameter	**Value**
Velocity (Speed) of SAT	$\mathbf{v}_{\mathrm{L,I}}^{1} = [0, 7590, 0]^T$ (7590 [m/s])
Radius of an orbit	$r_{\mathrm{E}} = 6921$ [km]
Orbital period	$T = 5728$ [s]
Number of SATs per orbital lane	$I = 22$
Inter-SAT distance	1977 [km]
Number of IoT nodes	$K = 10 - 1000$
Number of preamble signatures	$R = 64$

A. MDP Modeling

DRL framework for IoT-NTN is based on the environment where IoT nodes attempts to access to LEO SAT constellation. At each time n, a K set of IoT node, observes a state $\boldsymbol{s}[n]$ from the state space S, and accordingly takes an action $\boldsymbol{a}[n]$ from the action space A. Following the action, the state of the environment transitions to a new state $\boldsymbol{s}[n+1]$ and the centralized agent receives a reward $\boldsymbol{r}[n]$.

In the aforementioned MDP model, we consider the following state information:

$$\boldsymbol{s}[n] = \{n, \boldsymbol{q}_i[n], \sum_{k \in \mathcal{K}} c_k[n], \boldsymbol{a}[n-1]\}, \tag{7}$$

where $\boldsymbol{q}_{i_k}[n] \in \mathbb{R}^{I \times 3}, i \in \mathcal{I}$ denotes the position of SAT i, $c_j[n]$ corresponds an RA collision for this time slot. Note that the previous action $a[n-1]$ and current time slot n are used as a *fingerprint* to stabilize experience replay in the DRL.

The action space $\mathcal{A}$ in our environment is related to RA. Among SATs i, the agent UT j chooses one SAT to access by using the access action $a[n]$ in (2). The set of access action, $\mathcal{A}$, is defined as follows

$$\boldsymbol{a}[n] = \{w_1, w_2, \cdots, w_{|\mathcal{K}|}\}, \tag{8}$$

Our reward function is supposed to reinforce IoT nodes to carry out optimal access actions that minimize collision event while maximizing the access rate. It is worth noting that, the collision event occurs in stochastic, not in deterministic due to random preamble selection. This incurs too significant and frequent reward variation, hindering the training convergence.

Following the standard RL settings, we consider an environment in which RL agents interact for a given number of discrete time steps. At each time step n, the agent receives a state $\boldsymbol{s}[n]$ and selects an action $\boldsymbol{a}[n]$ from some set of possible actions A according to its policy π_θ, where π_θ is a mapping from states $\boldsymbol{s}[n]$ to actions $\boldsymbol{a}[n]$. In return, the agent receives the next state $\boldsymbol{s}[n+1]$ and receives a scalar reward $r[n]$. The process continues until the agent reaches a terminal state, after which the process restarts. The return $R[n] = \sum_{k=0}^{\infty} \gamma^k r[n+k]$ is the total accumulated return from time step n with discount factor $\gamma \in (0, 1]$. Following the principle of PPO [14], we train our agent which is omitted due to limit of the space. We note that the state of each IoT nodes is all locally observable information, while the reward function is centralized information that can be trained with centralized training and decentralized execution (CTDE).

IV. Numerical Evaluations

This section validates the proposed RA methods for IoT-NTN with respect to collision rate. For PPO, we identi-

Table II: Collision and success rate of our proposed PPO compared with baseline under $K = 10$ and $K = 100$ IoT nodes.

Scheme	Collision Rate C	Scheme	Collision Rate C
Heuristic ($K = 10$)	0.0341 (1.00x)	Heuristic ($K = 100$)	0.2781 (1.00x)
PPO ($K = 10$)	0.0079 (4.31x)	**PPO** ($K = 100$)	0.1379 (2.01x)

Table III: Normalized reward and average collision rate over training episodes under $K = 10$ and $K = 100$ IoT nodes.

Reward (# of IoT Nodes, K = 10)	Collision (K = 10)	Reward (K = 100)	Collision (K = 100)

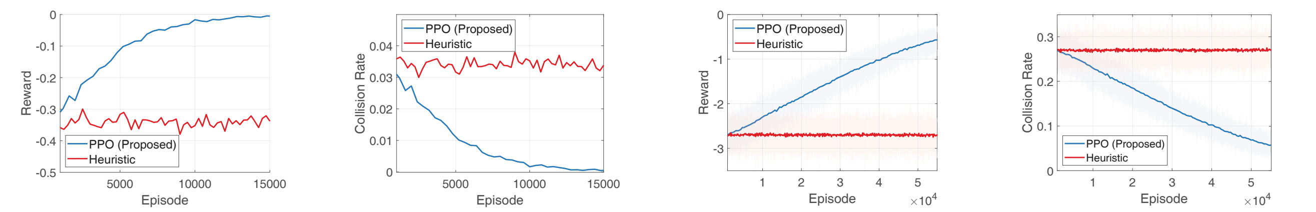

Table IV: Impact of the number of IoT nodes, K, on normalized reward and the number episodes until convergence.

# of IoT Nodes	Norm. Reward	# of Ep. for Convergence
$K = 10$	-0.008	$\simeq 1.5 \times 10^4$
$K = 100$	-0.138	$\simeq 5.5 \times 10^4$
$K = 1000$	-4.873	Diverge
	-5 -2.5 0	10^4 10^5 ∞

cally consider a 4-layer fully-connected multi-layer perceptron (MLP) NN architecture. Each MLP has 2 hidden layers, each of which has the 256×256 dimension with the rectified linear unit (ReLU) activation functions. NNs are trained using tanh with the learning rate 0.00005, batch size 4000, 10 iterations per episode, training episodes $10^4 \sim 10^5$, and 4000 iterations per update. The simulations are implemented using Pytorch Version 1.12. Main parameters on simulation settings are summarized in Table I.

Throughout this section, we consider one baseline and our proposed eRACH with as listed below.

1) **Heuristic** is a policy of random selection. Each IoT nodes uniform-randomly chooses access action, $w_k[n]$, at each access opportunity.
2) **PPO** is our proposed scheme with using PPO framework, wherein each IoT nodes determine its access action by following the policy of PPO which is trained in a fully centralized manner.

The following conditions are assumed in the comparison: if collisions occur at a time slot for an SAT, all attempted UTs fail to access at the time slot to the SAT.

Comparison Study. Table. II compare our proposed PPO with a baseline, in terms of collision rate and access rate. The results validate that PPO achieves lower collision with higher access rate than the baselines. In particular, compared to heuristic, PPO successfully access with 4.31x and 2.01x less collision event, for $K = 10$ and $K = 100$, respectively. As opposed to the heuristic method, PPO optimizes access policy, flexibly determining backoff action (i.e. $w_k[n] = 0$) for yielding higher access and lower collision in a given network scenario. Such flexibility is advantageous to access (or collision)-sensitive services such as massive machine type communications (mMTC).

Impact of the number of IoT nodes. Table III plot the normalized reward and the collision rate, while showing the training convergence behavior of PPO NN. Here, the solid curves denote cumulative reward for the agent. One can notice that more IoT nodes requires a more training episode for convergence of the DRL agent, as it incur a bigger action space and state space. It is worth noting that large action space is one of the main issue in the DRL approach, which is elaborated in the following.

Table IV show the impact of number of IoT nodes with respect to training convergence. In particular, around five times of training episodes is required for $K = 100$ compared to $K = 10$, while $K = 1000$ case is not converged within a given time. As the number of nodes increases, the size of action and state in DRL training increase depends on its training environment; and such increase causes additional training time, sometimes resulting the undesirable training behavior. For the large-scale systems, it is thus necessary to accelerate and stabilize the training convergence. In this regard, scalable DRL approach could be an interesting topic for our future research.

V. Conclusion

In this article, we proposed a novel RA for massive IoT in LEO SAT networks. To cope with the challenges incurred by its wide coverage and time-varying network topology, we proposed a model-free RA protocol based on the DRL approach. By simulations, we validated that our proposed PPO approach flexibly determines the access (or backoff) action over its given network scenario. Extending the current collision objectives, and considering other performance metrics (e.g., fairness-aware) could be an interesting topic for future research. It is also worth investigating highly scalable DRL frameworks to address complicated scenarios.

Acknowledgement

This research was funded by the National Research Foundation of Korea (2022R1A2C2004869). The part of this work was supported by the National Science Foundation of United States.

References

[1] Federal Communication Commission (FCC). (2018, Mar.) FCC authorizes SpaceX to provide broadband satellite services. [Online]. Available: https://www.fcc.gov/document/fcc-authorizes-spacex-provide-broadband-satellite-services

[2] ——. (2020, Jul.) FCC authorizes kuiper satellite constellation. [Online]. Available: https://www.fcc.gov/document/fcc-authorizes-kuiper-satellite-constellation

[3] Starlink, Accessed: Sep- 2021. [Online]. Available: https://www.starlink.com/

[4] N. Pachler, I. del Portillo, E. F. Crawley, and B. G. Cameron, "An updated comparison of four low earth orbit satellite constellation systems to provide global broadband," in *Proc. IEEE International Conf. on Commun. Workshops (ICC Wkshps)*, 2021, pp. 1–7.

[5] I. del Portillo, B. G. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," *Acta Astronautica*, vol. 159, pp. 123–135, 2019.

[6] 3GPP TR 36.763, "Study on narrow-Band internet of things (NB-IoT) / enhanced machine type communication (eMTC) support for Non-Terrestrial networks (NTN)," Jul. 2021.

[7] J.-H. Lee, K.-H. Park, Y.-C. Ko, and M.-S. Alouini, "Spectral-efficient network design for high-altitude platform station networks with mixed RF/FSO systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, Mar. 2022.

[8] J.-H. Lee, J. Park, M. Bennis, and Y.-C. Ko, "Integrating LEO satellite and UAV relaying via reinforcement learning for Non-Terrestrial networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, 2020, pp. 1–6.

[9] 3GPP TS 38.321, "NR; medium access control (MAC) protocol specification," Jul. 2021.

[10] 3GPP TS 38.213, "NR; physical layer procedures for control," Jun. 2021.

[11] 3GPP TR 21.916, "Release 16 description; summary of rel-16 work items," Sep. 2021.

[12] J.-H. Lee, H. Seo, J. Park, M. Bennis, and Y.-C. Ko, "Learning emergent random access protocol for LEO satellite networks," *IEEE Trans. Wireless Commun.*, pp. 1–14, May 2022.

[13] J.-H. Lee, H. Seo, J. Park, M. Bennis, Y.-C. Ko, and J. Kim, "Random access protocol learning in LEO satellite networks via reinforcement learning," 2022, pp. 1–5.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.