Multi-Agent Deep Reinforcement Learning For Distributed Handover Management In Dense MmWave Networks

Mohamed Sana, Antonio de Domenico, Emilio Calvanese Strinati, Antonio Clemente

MULTI-AGENT DEEP REINFORCEMENT LEARNING FOR DISTRIBUTED HANDOVER MANAGEMENT IN DENSE MMWAVE NETWORKS

Mohamed Sana, Antonio De Domenico, Emilio Calvanese Strinati, Antonio Clemente

CEA-Leti Minatec Campus, 17 rue des Martyrs, 38054 Grenoble Cedex 09, France

Email : {mohamed.sana, antonio.de-domenico, emilio.calvanese-strinati, antonio.clemente}@cea.fr

ABSTRACT

The dense deployment of millimeter wave small cells combined with directional beamforming is a promising solution to enhance the network capacity of the current generation of wireless communications. However, the reliability of millimeter wave communication links can be affected by severe pathloss, blockage, and deafness. As a result, mobile users are subject to frequent handoffs, which deteriorate the user throughput and the battery lifetime of mobile terminals. To tackle this problem, our paper proposes a deep multi-agent reinforcement learning framework for distributed handover management called **RHando** (Reinforced Handover). We model users as agents that learn how to perform handover to optimize the network throughput while taking into account the associated cost. The proposed solution is fully distributed, thus limiting signaling and computation overhead. Numerical results show that the proposed solution can provide higher throughput compared to conventional schemes while considerably limiting the frequency of the handovers.

Index Terms— Handover Management, mmWave, Multi-Agent Deep Reinforcement Learning

1. INTRODUCTION

The fifth generation (5G) of mobile communication systems will integrate millimeter-wave (mmWave) technologies to enable a significant improvement in the network capacity. In addition, network densification together with directional beamforming will allow to deal with the severe propagation losses that characterize mmWave communications [1]. Densification comes however with its own issues. As the number of base stations (BSs) increases, the handover (HO) rate of mobile user equipment (UEs) increases accordingly. In fact, a UE to maintain or improve its quality of service (QoS) may need to change its current BS association when moving through the network. Performing an HO procedure requires signaling between the UE, the serving BS, and the target BS, which induces overhead and energy consumption, thus decreasing the network performance. A wide range of HO algorithms exists in the literature, each trying to limit the impact of frequent HOs in UEs quality of experience (QoE). In general, HO decisions are based on measurement signals such as Received Signal Strength (RSS), Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), or Word Error Indicator (WEI) [2]. 3GPP standard suggests that a UE triggers an HO process when the RSS of the target BS exceeds the one of the serving BS by a certain amount to avoid ping pong effect [2]. This procedure may induce large signaling overhead preventing to meet the latency requirements of future wireless communication services [3]. To improve the HO performance, Yan *et al.* have proposed to limit the time-consumed in HO process by designing a machine learning algorithm that predicts HO decisions [4]. Koda *et al.* have proposed to limit HO frequency by designing a reinforcement learning (RL) framework that uses a Q-learning algorithm to maximize the network throughput [5]. In the same vein, Wang *et al.* have extended this approach using deep RL with actor-critic methods to avoid state discretization and for better scalabilty [6]. All these works do not consider cell load and limited resource availability when optimizing the HO strategy.

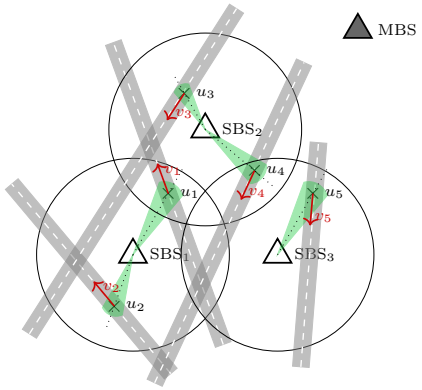


Fig. 1. A downlink network with $N_s = 3$ SBSs, one MBS, and $K = 5$ UEs. As an example, $\mathcal{U}_1(t) = \{1, 2\}$, $\mathcal{S}_5(t) = \{3\}$.

In this paper, we propose a distributed algorithm called RHando - Reinforced Handover - that optimizes handover decisions in order to maximize the network throughput, considering mobility and limited resources at the BS side. We model each UE as an agent, which at each HO triggering time, based on its local observation takes the handover decision. In this framework, each agent maintains its own deep recurrent Q network (DQRN) and decides regardless of other UEs. This limits both computational and signaling complexity since communications is not required between UEs, thus making RHando a good candidate to meet the latency requirements of 5G networks.

The rest of the paper is organized as follows. Section 2 describes the system model and formulate the handover decision problem. In Section 3 we present the proposed solution and we show numerical results in Section 4. Finally, Section 5 concludes the paper.

2. SYSTEM MODEL AND PROBLEM FORMULATION

2.1. System Model

In the considered downlink network deployment (see Fig. 1), K UEs are moving in an area where N_s millimeter waves small cell BSs (SBSs) are collocated with a sub-6 GHz macro base station (MBS) to provide broadband wireless services. We denote by $\mathcal{S}$ the set of

$N_s + 1$ BSs in the network and let $\mathcal{U}$ define the set of K UEs. Due to mobility, given a BS i, the set of UEs in its coverage area $\mathcal{U}_i(t)$ changes over time as well as the set of BSs a UE j could associate with, $\mathcal{S}_j(t)$. In our model, each SBS allocate all the mmWave band to each of its served UEs, which are spatially multiplexed through direct beamforming. In contrast, the MBS equally share its frequency resources across its UEs. In addition, we assume that a SBS can support at most N_i UEs at the same time.

To characterize the environment dynamic, we adopt a classic distance-based path loss model [7], where the channel fading coefficient follows a Nakagami-m distribution with a shape factor m [8]. Thereupon, the downlink rate perceived from the BS i by the UE j is as follows:

$$R_{i,j}(t) = x_{i,i}(t) B_{i,j} \log_2(1 + \text{SINR}_{i,j}(t)), \tag{1}$$

where $x_{i,i}(t) = 1$ if the UE j is associated with the BS i and $x_{i,i}(t) = 0$ otherwise. $\text{SINR}_{i,j}(t)$ denotes the signal-to-interference plus noise ratio, which includes both intra-cell and inter-cell interference.

2.2. Handover Overhead and Network Sum-rate Maximization

As UEs are moving across the network, they may be subject to multiple handovers in order to maintain or improve their QoE. However, unnecessary HOs lead to large signaling overhead, which increases the energy consumption, lowers the spectral efficiency, and affects UEs latency. To account with this, we directly introduce a penalty due to the handover in the evaluation of the network performance. Indeed, let $\Delta\tau$ be the time between two possible handovers, also known as Time-to-Trigger (TTT) interval [2]. That is, a handover process can be triggered every time $\tau_p = \tau_0 + p\Delta\tau$, where τ_0 is an initial system delay. If UE j want to perform a handover at time τ_p, then, a time $\beta\Delta\tau$ is dedicated to the handoff procedure while the time $(1-\beta)\Delta\tau$ is used to communicate data. The coefficient $\beta \in [0, 1]$ allows to control the cost of an HO process, which depends on the type of implemented handover (soft or hard handover) [9]. Accordingly, the effective data received by UE j from BS i between time τ_p and τ_{p+1} is

$$\overline{R}_{i,j}(\tau_p, \beta) = \int_{\tau_p}^{\tau_p + (1-\beta\lambda_j(\tau_p))\Delta\tau} R_{i,j}(t)dt, \tag{2}$$

where $\lambda_j(\tau_p) = 1$ indicates that UE j has handed over at time τ_p, and $\lambda_j(\tau_p) = 0$ otherwise. Hence, we define the network throughput $R(\tau_p)$ measured between time τ_p and τ_{p+1} as follows:

$$R(\tau_p, \beta) = \frac{1}{\Delta\tau} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{U}} \overline{R}_{i,j}(\tau_p, \beta). \tag{3}$$

Let $\lfloor . \rfloor$ be the floor operator and $P = \lfloor T/\Delta\tau \rfloor$ be the number of TTTs over a time period T. We aim to find the HO strategy that maximizes the average network throughput $R_T(\beta) = \frac{1}{T}\sum_{p=1}^{P} R(\tau_p, \beta)$ taking into account the cost associated to handoff events. Hence, we formalize this problem as follows:

$$\max_{\{x_{i,j}\}} \quad R_T(\beta) \tag{4}$$

$$\text{s.t.} \quad x_{i,j}(\tau_p) \in \{0, 1\}, \qquad i \in \mathcal{S},\ j \in \mathcal{U},\ p = [1, P], \tag{5}$$

$$\sum_{j \in \mathcal{U}_i} x_{i,j}(\tau_p) \leq N_i, \qquad i \in \mathcal{S}\backslash\{0\},\ p = [1, P], \tag{6}$$

$$\sum_{i \in \mathcal{S}_j} x_{i,j}(\tau_p) = 1, \qquad j \in \mathcal{U},\ p = [1, P]. \tag{7}$$

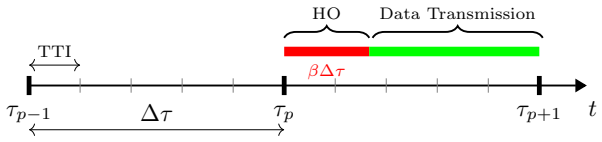


Fig. 2. HO process timeline. TTI denotes the Transmission Time Interval.

Constraint (5) ensures that the decision variables are binary. Constraint (6) indicates that the maximum number of UEs that a SBS can simultaneously support is limited to N_i. Finally, constraint (7) indicates that a UE is always associated with a BS. The optimization problem (4)-(7) is a non-convex integer programming problem. In addition to the complexity of such a problem, the optimal association at time τ_p also depends on the association at time step τ_{p-1} through the handover variable λ_j, making the problem (4)-(7) intractable with conventional optimization frameworks. In the following, we propose a solution based on multi-agent reinforcement learning to tackle this problem.

3. PROPOSED HANDOVER FRAMEWORK

In this section, we depict the proposed HO solution. We formalize the optimization problem (4)-(7) as a multi-agent reinforcement learning (MARL) task where each UE is modeled as an independent agent that learns in a distributed way its handover strategy with the goal of optimizing the network throughput.

3.1. General Setup

3.1.1. ***UEs action space***

At each time step τ_p, each UE j takes an action $a_j(\tau_p)$ to associate with one BS in the network. If the connection request is addressed to the MBS, this is automatically granted. Otherwise, if the requested SBS is able to support the association, an acknowledgment signal is sent (ACK=1), otherwise ACK=0 (see constraint (6)). Finally, if the BS that UE j is effectively associated with at time step τ_p differs from the one at time step τ_{p-1}, the UE initiates an handover procedure. Later, the MBS collects information from each BS to compute the overall network throughput $R(\tau_p, \beta)$, which is broadcast to all UEs to evaluate the goodness of their policy.

3.1.2. ***UEs state space***

To learn their optimal strategy, UEs continuously collect information about their surrounding environment. We assume that at each time step, each UE can measure the RSS of the surrounding BSs i.e., $\{\text{RSS}_i, \forall i \in \mathcal{S}\}$. In addition, each UE uses the previously perceived data rate $R_{a_j(\tau_p),j}(\tau_{p-1}, \beta)$ and network sum-rate $R(\tau_{p-1}, \beta)$. Hence, at time τ_p, UE j acts based on local observations:

$$\begin{aligned} o_j(\tau_p) \quad = \big\{ & v_j^x(\tau_p), v_j^y(\tau_p), a_j(\tau_{p-1}), \overline{R}_j(\tau_{p-1}, \beta), \\ & R(\tau_{p-1}, \beta), \text{ACK}_j(\tau_{p-1}), \{\text{RSS}_i(\tau_p)\}_{\forall i \in \mathcal{S}} \big\}, \end{aligned} \tag{8}$$

where $v_j(\tau_p) = (v_j^x(\tau_p), v_j^y(\tau_p))$ is the corresponding UE's speed.

3.1.3. ***UEs reward***

To optimize the network performance, UEs must learn how to perform association requests that limit handovers and avoid *collisions* across service requests.

Definition 3.1 *Let $c(\tau_p)$ denotes the request collision event. There is a request collision at time step t, i.e., $c(\tau_p) = 1$, if $\exists$ i such that $\sum_{j\in\mathcal{U}} x_{i,j}(\tau_p) > N_i$. Otherwise, $c(\tau_p) = 0$.*

Hence, collisions occur when the number of UEs requesting an association with a given SBS is greater than the number of connections (i.e., N_i) the SBS can support. To optimize the system, we have designed two reward functions taking into account the collision events.

- **RHando-F (Fully cooperative RHando)**: in this strategy, UEs receive the same reward, which favors global network optimization:

$$r_j(\tau_p) = (1 - c(\tau_p))\Delta\tau R(\tau_p, \beta). \tag{9}$$

- **RHando-S (Self interest RHando)**: here, each UE instantaneous reward only considers the data rate it perceived. Hence,

$$r_j(\tau_p) = (1 - c(\tau_p))\overline{R}_{a_j(\tau_p),j}(\tau_p, \beta). \tag{10}$$

  It is noteworthy that even in RHando-S, the reward of each UE still depends on other UEs because of the interference and the collision events.

Next, each UE j learns a policy that maximizes its long term γ-discounted reward:

$$G_j = \sum_{p=1}^{P} \gamma^{\tau_{p-1}} r_j(\tau_p). \tag{11}$$

3.2. Learning procedure

The agents learn their handover strategy with the caveat of hysteretic DRQN [10], which enables them to leverage on aggregated past information. At each time step, UE j observes $o_j(\tau_p)$, takes an action $a_j(\tau_p)$ that brings it to a new state $o_j(\tau_{p+1})$. As a consequence of that action, the UE receives an immediate reward $r_j(\tau_p)$ (defined as (9) or (10)). Then, the resulting experience $e_j(\tau_p) = (o_j(\tau_p), a_j(\tau_p), r_j(\tau_p), o_j(\tau_{p+1}))$ is stored into a local memory $\mathcal{M}_j$. During the training process, all the agents synchronously sample a batch of experiences from their local memory using concurrent experience replay [10]. Then, each agent j updates its DRQN weights by minimizing the loss function:

$$L_j(\theta_j) = \mathbb{E}_{e_j^b(\tau_p)\sim\mathcal{B}_j}\left[\left(w_j^b \delta_j^b(\tau_p)\right)^2\right]. \tag{12}$$

In (12), b indexes an entry in the mini batch of experiences $\mathcal{B}_j$. $\delta_j^b(\tau_p) = y_j^b(\tau_p) - Q_j(o_j^b(\tau_p), h_j^b(\tau_{p-1}), a_j^b(\tau_p)|\theta_j)$ denotes the temporal difference (TD) error between the output of the UE j DRQN and the target value:

$$y_j^b(\tau_p) = r_j^b(\tau_p) + \gamma \max_{a'} Q_j(o_j^b(\tau_{p+1}), h_j^b(\tau_p), a'|\hat{\theta}_j), \tag{13}$$

where $\hat{\theta}_j$ are the weights of the target DRQN. In the hysteretic Q-learning, the neural network weights are updated via a gradient descent algorithm with two distinct learning rates $\zeta\mu$ and $\eta\mu$ ($\eta \ll \zeta \leq 1$) [10], where μ is a fixed base learning rate. When the TD error $\delta \geq 0$, the learning rate $\zeta\mu$ is used; otherwise, $\eta\mu$ is considered. This leads to an optimistic update that gives more importance to positive experiences [11]. Accordingly, the weights w_j^b in (12) are defined as follow:

$$w_j^b = \begin{cases} \zeta, & \text{if } \delta_j^b(t) \geq 0 \\ \eta, & \text{otherwise} \end{cases}. \tag{14}$$

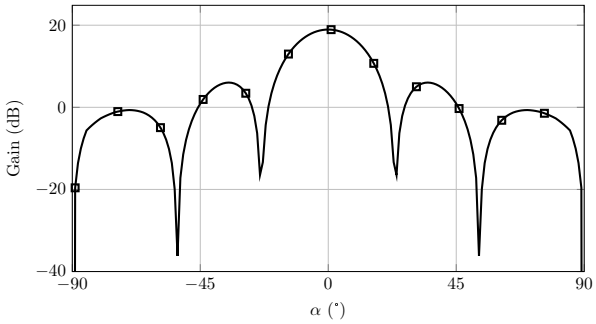


Fig. 3. Simulated TX/RX antenna gain radiation pattern for an array of 5×5 elements operating at 28 GHz [12].

Table 1. Simulation parameters.

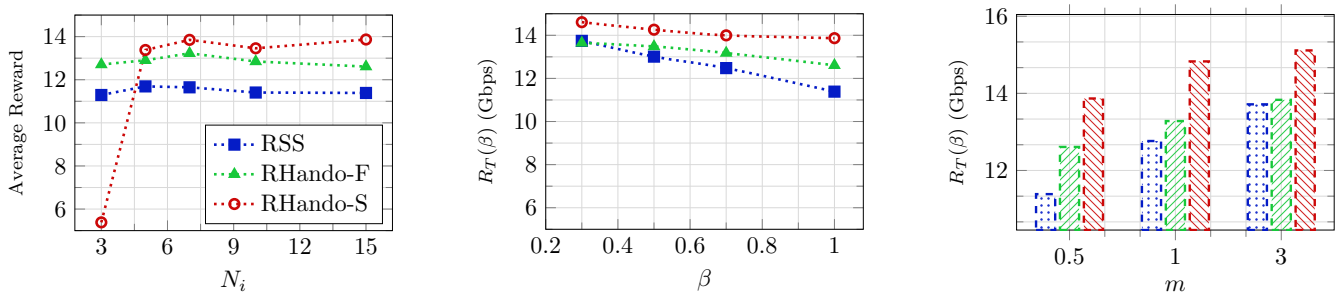
	Macro cell [13]	**Small cell**
Parameters	Values	
Carrier frequency, f_c	2.0 GHz	28 GHz
Bandwidth, B	10 MHz	500 MHz
Thermal Noise, N_0	-174 dBm/Hz	
Noise figure	5 dB	0 dB
Shadowing, X	9 dB	12 dB
Transmit power	46 dBm	20 dBm
Antenna gain (TX/RX)	17 dBi / 0 dBi	Fig. 3
Cell radius, r		35 m
Inter-cell distance		$1.2 \times r$
Pathloss model	$128.1 + 36.7\log_{10}(d)$	[7], $d_0 = 5$ m
TTI	10 ms	
$\Delta\tau$	1 s	
T	2000 s	

4. NUMERICAL RESULTS

To assess the performance of the proposed framework, we consider as a benchmark a simplified version of the HO procedure proposed in 3GPP [2] in which each UE is associated to the BS providing the strongest RSS. In case of request collision, each SBS selects the best UEs in terms of RSS while the MBS serves the others UEs.

In all tests, five mmWave SBSs are deployed inside the macro cell. UEs locations are randomly initialized. To account for heterogeneous mobility, each UE randomly picks a speed between 0 and 10 ms^{-1} and takes a straight motion with random direction. In addition, without loss of generality, we suppose that users turn back once they reach the macro cell edge. Finally, each UE is equipped with a dueling DRQN (with averaged advantages)[14]. This architecture comprises two multi-layers perceptron (MLP) of 32 hidden units, one long short term memory (LSTM) layer with 64 memory cells, followed by another two MLPs of 32 hidden units. The network then branches off in two MLPs of 16 hidden units [14]. We fix the discounting factor $\gamma = 0.9$, the base learning rate $\mu = 0.001$ and empirically set $\eta = 0$ and $\zeta = 0.4$ through exhaustive search. Finally, we train the DQRNs using an ϵ-greedy policy with ϵ annealing from 1 to 0.1.

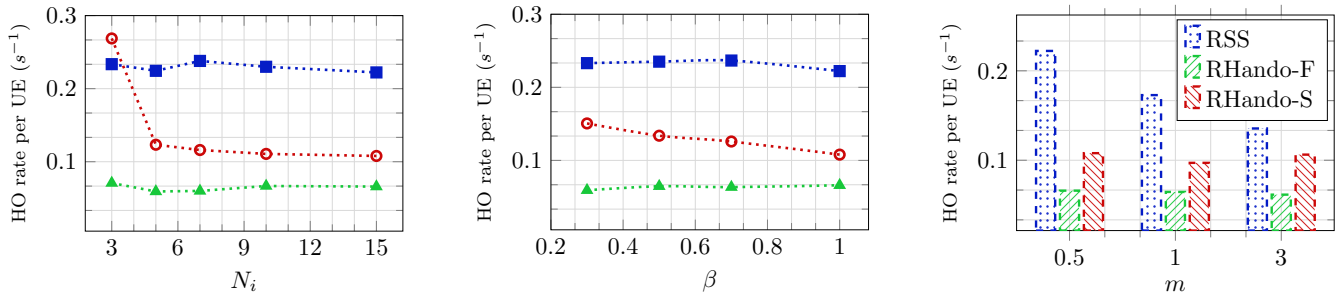
For a given UE i associated to a given BS j, we evaluate $\overline{R}_{i,j}(\tau_p, \beta)$ by aggregating the data received during each TTI (see Eq. (2)). Fig. 3 shows the transmitter and receiver beamforming gains in the mmWave band. Additional simulation parameters can be found in Table 1.



(a) Avg. reward w.r.t. N_i, $K = 15$, $m = 0.5$, $\beta = 1$.

(b) $R_T(\beta)$ w.r.t. β. $N_i = K = 15$, $m = 0.5$.

(c) $R_T(\beta)$ w.r.t. m. $N_i = K = 15$, $\beta = 1$.



(d) HO rate w.r.t. N_i. $K = 15$, $m = 0.5$, $\beta = 1$.

(e) HO rate w.r.t. β. $N_i = K = 15$, $m = 0.5$.

(f) HO rate w.r.t m. $N_i = K = 15$, $\beta = 1$.

Fig. 4. Performance comparison w.r.t. number of possible beams, HO cost factor, and Nakagami scale factor.

4.1. Performance in terms of collision avoidance

As mentioned in Subsection 3.1.3, request collisions may happen when BSs do not have enough beams to simultaneously support all the service requests. Figs. 4a and 4d show the performance comparison of the two RHando configurations compared to the benchmark solution. Unsurprisingly, for lower values of N_i, RHando-S exhibits poor performance than RHando-F both in terms of average reward (i.e., $(1/P)\sum_{p=1}^{P}\sum_{j\in\mathcal{S}} r_j(\tau_p)$) and HO's frequency. This is because UEs in RHando-F fully cooperate through the common reward they perceive and, as a result, they effectively learn to avoid request collisions. In contrast, with RHando-S, each UE learns a policy based on a local reward, which does not provide sufficient information on the effect of its action on the other UEs' reward. Inversely, when N_i is sufficiently large (> 7), RHando-S outperforms both RHando-F and RSS-based HO in terms of average reward. The throughput is increased by about 17.89% by RHando-S and only 10% by RHando-F compared to the benchmark. Regarding the HO events, RHando-F decreases the HO frequency by about 70% and RHando-S by 54% compared to the baseline. Overall, we can observe that the fully cooperative approach limits the handover rate at the cost of lower reward when N_i is large.

4.2. Performance w.r.t. handover cost factor β

Now we evaluate the performance of the proposed solutions with respect to the handover cost factor β. Fig. 4b shows that when the HO cost increases, the network average throughput decreases. The RSS-based solution is characterized by the worst performance as it does not consider the handover cost. Fig. 4e shows that when the HO becomes more and more costly, the HO rate decreases with Rhando-S while remaining almost constant with Rhando-F. This is because the HO cost variation has a limited impact on the global reward perceived by the agents in RHando-F: after a handoff decision, an agent can still perceive a large global reward as this is defined as the sum of all the other agents' rewards.

4.3. Average throughput w.r.t. Nakagami fading scale factor m

HO events highly depends on the channel conditions viz. path loss and fading. Here, we evaluate the performance gain of the different algorithms with respect to Nakagami scale factor m. Figs. 4c and 4f show that the more severe the fading ($m \to 0$), the more pronounced the gain of the proposed solution compared to the benchmark both in term of average throughput and number of HOs. The performance of the RSS-based HO strongly deteriorates with the fading while RHando-F and RHando-S adapt their policies to the fading characteristics demonstrating therefore the robustness of the proposed framework.

5. CONCLUSION

In this work, we have proposed a framework to manage handover events based on multi-agent deep reinforcement learning. We maximize the average network sum-rate taking into account the cost induced by HO events. The proposed solution is distributed among UEs thus limiting signaling overhead. We show that it can reduce the HO events up to 70% and increase the average network throughput by 18% compared to a HO solution based on maximum RSS.

Future work will focus on heterogeneous access network integrating HO strategies based on multi-connectivity.

6. REFERENCES

[1] A. De Domenico, R. Gerzaguet, N. Cassiau, A. Clemente, R. D'Errico, C. Dehos, J. L. Gonzalez, D. Ktenas, L. Manat, V. Savin, and A. Siligaris, "Making 5G Millimeter-Wave Communications a Reality," *IEEE Wireless Communications*, vol. 24, no. 4, pp. 4–9, Aug 2017.

[2] 3GPP TR 36.839 V11.1.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility enhancements in heterogeneous networks (Release 11)," Jan 2013.

[3] 3GPP TS 22.261 V17.0.1, "Service Requirements for Next Generation New Services and Markets (Release 17)," Oct 2019.

[4] L. Yan, H. Ding, L. Zhang, J. Liu, X. Fang, Y. Fang, M. Xiao, and X. Huang, "Machine Learning Based Handovers for Sub-6 GHz and mmWave Integrated Vehicular Networks," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2019.

[5] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Reinforcement learning based Predictive Handover for Pedestrian-aware mmWave Networks," in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr 2018, pp. 692–697.

[6] Z. Wang, L. Li, Y. Xu, H. Tian, and S. Cui, "Handover Optimization via Asynchronous Multi-User Deep Reinforcement Learning," in *Proc. IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[7] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Mobility Management for Heterogeneous Networks: Leveraging Millimeter Wave for Seamless Handover," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec 2017, pp. 1–6.

[8] R. Chevillon, G. Andrieux, R. Négrier, and J. Diouris, "Spectral and Energy Efficiency Analysis of mmWave Communications With Channel Inversion in Outband D2D Network," *IEEE Access*, vol. 6, pp. 72104–72116, 2018.

[9] H. Park, Y. Lee, T. Kim, B. Kim, and J. Lee, "Handover Mechanism in NR for Ultra-Reliable Low-Latency Communications," *IEEE Network*, vol. 32, no. 2, pp. 41–47, Mar 2018.

[10] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability," in *Proc. International Conference on Machine Learning (PMLR)*, 06–11 Aug 2017, vol. 70, pp. 2681–2690.

[11] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Hysteretic Q-learning: an Algorithm for Decentralized Reinforcement Learning in Cooperative Multi-agent Teams," in *Proc. International Conference on Intelligent Robots and Systems (IEEE/RSJ)*, 2007, pp. 64–69.

[12] Robert J. Mailloux, *Phased Array Antenna Handbook*, Artech House, Inc., Norwood, MA, USA, 3rd edition, 2017.

[13] 3GPP TR 36.872 V12.1.0, "Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects (Release 12)," Dec 2013.

[14] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling Network Architectures for Deep Reinforcement Learning," in *Proc. International Conference on Machine Learning (PMLR)*, Jun 2016, vol. 48, pp. 1995–2003.