2024 EuCNC & 6G Summit - WOS 1570990245
2024 European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit): Wireless, Optical and Satellite Networks (WOS)

Decentralized RL-Based Data Transmission Scheme for Energy Efficient Harvesting

Rafaela Scaciota[1], Glauber Brante[2], Richard Souza[3], Onel Lopez[1],
Septimia Sarbu[1], Mehdi Bennis[1], and Sumudu Samarakoon[1]
[1]*Centre for Wireless Communication, University of Oulu, Finland*
[2] *Federal University of Technology - Paraná, Brazil*
[3]*Federal University of Santa Catarina, Brazil*

***Abstract*—The evolving landscape of the Internet of Things (IoT) has given rise to a pressing need for an efficient communication scheme. As the IoT user ecosystem continues to expand, traditional communication protocols grapple with substantial challenges in meeting its burgeoning demands, including energy consumption, scalability, data management, and interference. In response to this, the integration of wireless power transfer and data transmission has emerged as a promising solution. This paper considers an energy harvesting (EH)-oriented data transmission scheme, where a set of users are charged by their own multi-antenna power beacon (PB) and subsequently transmits data to a base station (BS) using an irregular slotted aloha (IRSA) channel access protocol. We propose a closed-form expression to model energy consumption for the present scheme, employing average channel state information (A-CSI) beamforming in the wireless power channel. Subsequently, we employ the reinforcement learning (RL) methodology, wherein every user functions as an agent tasked with the goal of uncovering their most effective strategy for replicating transmissions. This strategy is devised while factoring in their energy constraints and the maximum number of packets they need to transmit. Our results underscore the viability of this solution, particularly when the PB can be strategically positioned to ensure a strong line-of-sight connection with the user, highlighting the potential benefits of optimal deployment.**

***Index Terms*—Wireless Powered Systems, Energy Harvesting, Power Beacon, Irregular Slotted Aloha, CSI.**

I. Introduction

By the end of 2023, it is projected that Internet of Things (IoT) users will constitute approximately 50% of all networked users worldwide, as reported in [1]. In this dynamic and swiftly evolving realm of the IoT, the need for efficient and dependable communication methods has reached unprecedented importance [2]. As this expansive and continually growing ecosystem of IoT users continues to scale up, conventional communication protocols face substantial challenges in meeting the requirements of this burgeoning network. Within this context, the integration of wireless power transfer and wireless data transfer has emerged as a pivotal solution, opening up a new era of possibilities for IoT users [3].

In this context, works like [4], [5] have demonstrated the feasibility of harvesting radio frequency (RF)-energy from sources such as a power beacon (PB). Wireless-powered systems face the challenge of efficiently utilizing harvested energy while minimizing packet collisions for users. To address this issue, simple Aloha protocols are commonly employed. These protocols are favored for their signaling simplicity and energy efficiency at the transmitter, which ultimately helps reduce collisions and optimize energy utilization. With this aim, prior works consider the use of irregular slotted aloha (IRSA) protocol for powering the users over RF. In [6], the authors present a feedback-aided IRSA scheme that improves the user energy efficiency by optimizing the transmit power and the number of packet replicas using high-throughput transmission probability distributions. With the focus on improving wireless powered systems, the authors in [7] further extend the discussion by proposing an IRSA-based uncoordinated random access scheme for EH nodes. It is considered a scheme where each user has a battery that is recharged by an EH system. The results showcase optimized probability distributions for packet replicas and highlight the improvement of performance in IRSA protocol. In [8], the authors introduce an IRSA protocol tailored for resource-constrained nodes in wireless energy transfer environments. Therein, the concept of a hybrid access point (HAP) is used and the optimal threshold value that maximizes throughput in these unique networks is investigated.

A learning-based solution for addressing communication protocols integrated with wireless power transfer challenges can be found in [9] where the authors delve into the integration of IRSA with RF users, with a particular emphasis on the critical task of optimizing the number of packet replicas. What sets this research apart from previous approaches is the utilization of a *Q*-learning-based methodology. This approach allows for the dynamic adjustment of the number of replicas based on the energy levels of the users, resulting in substantial improvements in the success rate of transmissions. However, it is worth noting that certain gaps related to the utilization

This work has been partially supported in Brazil by CAPES, Finance Code 001, CNPq (402378/2021-0, 305021/2021-4, 307226/2021-2). This work is funded by the European Union Project CENTRICunder Grant Agreement (GA 101096379), VERGE (GA 101096034) the project Infotech R2D2, the Research Council of Finland (former Academy of Finland) (GArant n. 348515), and the Finnish Foundation for Technology Promotion. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission (granting authority). Neither the European Union nor the granting authority can be held responsible for them. Corresponding author: rafaela.scaciotasimoesdasilva@oulu.fi

of the IRSA protocol in wireless-powered systems, such as scalability, interference, latency, and hardware cost, still exist in the literature. These gaps represent opportunities for further research and exploration in this evolving field.

Inspired by the existing works, we present a novel joint data transmission and EH for IoT networks. In this scheme, individual users are powered by their respective PB as a distributed EH, employing average channel state information (A-CSI) beamforming techniques. We compare the use of A-CSI with full channel state information (F-CSI) technique which assumes the availability of perfect instantaneous F-CSI for the user-PB link. Subsequently, each user proceeds to transmit data to a central BS using an IRSA channel access protocol. Our paper provides a comprehensive closed-form expression that accurately models energy consumption within this framework. From an EH perspective, the energy beamforming scheme empowers a multi-antenna power beacon to efficiently deliver power to individual users, solely relying on the first-order statistics of channel conditions. This approach effectively mitigates interference from other user-PB systems within the setup, thanks to the integration of distributed EH.

Moreover, to address the EH-oriented data transmission scheme, we adopt an approach based on reinforcement learning (RL). In this context, each user acts as an agent with the objective of discovering their optimal strategy for transmitting replicas, taking into consideration their energy limitations and the maximum number of packets to be transmitted. We leverage independent *Q*-learning to showcase the scalability of our system. The scalability allows for an increase in the number of users, providing practical feasibility for addressing the challenges in EH-oriented data transmission schemes. The significance of our learning methodology becomes evident when we draw comparisons with the contention resolution diversity slotted Aloha (CRDSA) channel access protocol, which consistently sends the same number of packets per user [10]. Our numerical results further emphasize the positive impact of incorporating the *Q*-learning method, demonstrating about 18% increase in successful transmissions per frame compared to a baseline scheme employing the CRDSA channel protocol without any learning process.

II. System Model

We assume the scenario illustrated in Fig. 1 where a set of U single-antenna users are charged by personal (or own) PBs. Each PB is equipped with a uniform linear array (ULA) of M antennas. The users need to harvest RF energy from PBs to send data to the base station (BS), which results in a waiting period referred to as the charging slot. The PB transmits at a fixed transmit power during the charging slot. The user-BS communications use IRSA channel access protocol where K data slots $\{\mathrm{d}_k\}_{k=1}^{K}$ are allocated. Time is discrete and indexed by t.

A. Irregular slotted aloha (IRSA)

In the context of IRSA for data transmission, a user employs multiple replicas of a packet within each frame as shown in Fig. 2a. These packets contain information about the slots in which other replicas are sent. The transmission of a packet and its replicas can be represented as a bipartite graph, as depicted in Fig. 2b. Each user is represented by a circle, while each data slot is represented by a square. The edges connect a user to the selected data slots. For an example, user u_1 has chosen to transmit in slots d_1 and d_2, whereas user u_2 transmits in slots d_1, d_2, and d_3, while user u_3 transmits only in slot d_2.

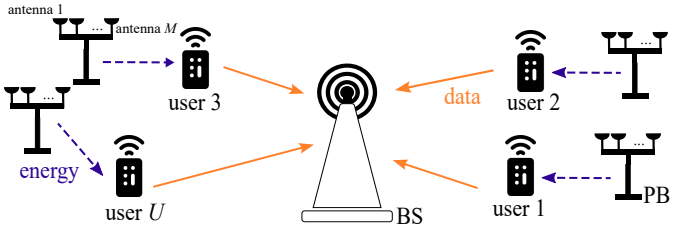

Figure 1: Wireless power transfer (WPT) with a BS and U user's RF-EH system model.

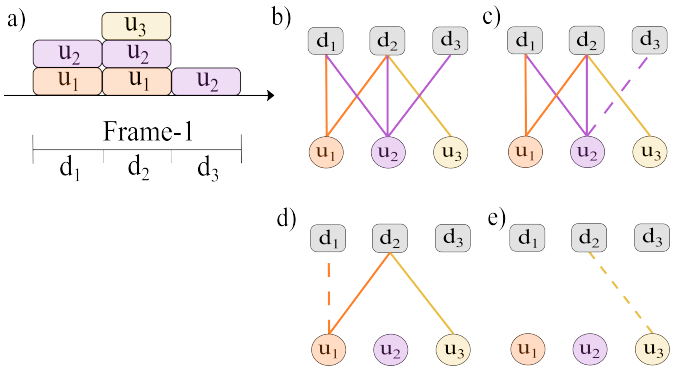

Figure 2: Bipartite representation of IRSA with its interference cancellation mechanism. (a) Frame structure. (b) Bipartite graph. (c) Iteration 1. (d) Iteration 2. (e) Iteration 3.

Once the replicas are received, the BS decodes the packets starting from a data slot with no collisions (Fig. 2c) as follows [11]. The decoded packet serves as a reference to decode packets in other data slots in which, all copies of the decoded packet from the corresponding user are removed first. Then, the BS seeks another collision-free data slot to decode the next packet (Fig. 2d). This is repeated (Fig. 2e) until all data from users are sequentially decoded.

We assume that user u sends $\psi \in \{0, \ldots, N\}$ replicas of a given packet in time frame t with probability π_ψ. The probability mass function that determines the probability of sending a given number of replicas is represented as the polynomial $\vec{\pi}(x) = \sum_{\psi=0}^{N} \pi_\psi x^\psi$.

B. Channel Model

We assume users and the BS are equipped with a single antenna and the channel between a user and the BS (used for data transmission) is subject to Rayleigh fading. In addition, we assume a frequency hopping model, where transmission occurs in different channel realizations. Therefore, $g_u = \sqrt{1/2}\mathcal{CN}(\mathbf{0}, PL)$ is modeled as a complex Gaussian random

variable with zero mean and $PL = 20\log_{10}(4c\pi d_u^{\text{BS}}/f)$ variance where c is the speed of light, f is the carrier frequency, and d_u^{BS} is the BS-user distance.

In EH, the channel between the PB and each user is subjected to quasi-static Rician fading due to the line-of-sight (LOS) connectivity. Thus, the channel vector between each user and the PB is given by [12]

$$\mathbf{h}_u = \sqrt{\beta_u}\left(\overline{\mathbf{h}}_u + \tilde{\mathbf{h}}\right) \quad \in \mathbb{C}^{M\times 1}, \tag{1}$$

where $\beta_u = c^2/(16\pi^2 f^2 (d_u^{\text{PB}})^\alpha)$ is the average power gain, c is the speed of light, f is the carrier frequency, d_u^{PB} is the PB-user distance, and α is the pathloss exponent. $\overline{\mathbf{h}}_u = \sqrt{\kappa/(2(1+\kappa))}\,[1, e^{i\tau_1}, \cdots, e^{i\tau_{M-1}}]^T$ is the deterministic LOS component, and $\tilde{\mathbf{h}} \sim \sqrt{1/(1+\kappa)}\mathcal{CN}(\mathbf{0}, \mathbf{R})$ is the zero-mean scattering (random) component with covariance $\mathbf{R} = \mathbb{E}\left[\tilde{\mathbf{h}}\tilde{\mathbf{h}}^H\right]$ where τ_m, $m \in \{1, \cdots, M-1\}$, is the mean phase shift of the $(m+1)$-th array element with respect to the first antenna and κ is the LOS factor [13]. Assuming half-wavelength spacing between antenna elements $\tau_m = -m\,\pi \sin\theta$ is held, where $\theta \in [0, 2\pi]$ is the azimuth angle relative to the boresight of the transmitting antenna array.

C. Incident Power

Accounted for a single user, the maximum ratio transmission (MRT) precoding is used as the optimal precoder design. Hence, the exact channel $\mathbf{h}_u$ is used for the precoder with F-CSI while for A-CSI, the expected channel $\mathbb{E}[\mathbf{h}_u] = \sqrt{\beta_u}\,\overline{\mathbf{h}_u}$ is used. In this view, the harvested power is given by [14],

$$P_U^i = \begin{cases} \beta_u P_{\text{b}} ||\mathbf{h}_u||^2, & i = \text{F-CSI}, \\ \beta_u\, P_{\text{b}} \left| ||\overline{\mathbf{h}_u}|| + \frac{\overline{\mathbf{h}_u}^H \tilde{\mathbf{h}}}{||\overline{\mathbf{h}_u}||} \right|^2, & i = \text{A-CSI}. \end{cases} \tag{2}$$

D. Energy Model

Assuming that user u sends replicas ψ_u^t in the time frame t, the energy level of user u evolves as per

$$E_u^t = \min\left(E_0,\, E_u^{t-1} + G(P_u)t_{\text{C}} - \frac{(1+\rho_u^t)\xi_u |g_u|^2}{(d_u^{\text{BS}})^\alpha}\right) \tag{3}$$

where $\xi_u = t_{\text{T}} P_u L$ is the energy spent in the transmission of one replica, L is the packet size, P_u is the transmission power, t_{C} is the charging slot time, t_{T} is the data slot time, and $G(P_u) = \mathcal{W}(1 - e^{-c_0 G(P_u)})/(1 + e^{-c_0(G(P_u) - c_1)})$ is a non-linear EH function [15], [16], with $\mathcal{W}$ being the saturation level, and c_0 and c_1 are constants. The battery capacity is $E_0 = \omega\xi_u$, corresponding to the energy required to send ω packets. We define energy levels as discrete. Note that, the maximum number of replicas user u can send at time t is $l_u^{\max} = \frac{E_u^t}{\xi_u} - 1$.

III. Packet Decode Maximization Problem

Our objective is to identify the optimal policy, denoted as ϕ, that maximizes the successful packet decoding throughout a planning horizon of duration H. In this context, each user policy $\phi_u = \{\boldsymbol{o}_u^1, \boldsymbol{o}_u^2, \ldots, \boldsymbol{o}_u^H\}$ dictates the allocation of replicas to each user during individual time frames, where $\boldsymbol{o}_u^t$ is a vector of size ψ. We establish a collection of optimal policies denoted as $\Phi = \{\phi_1, \phi_2, \ldots, \phi_u\}$. The quantity of successfully decoded messages during time frame t is represented as $\mathcal{R}_u^t$. Formally, our problem is formulated as follows:

$$\mathcal{R}_u = \max_{\phi_u \in \Phi} \frac{1}{H}\,\mathbb{E}\left[\sum_{t=1}^{\infty} \mathcal{R}_u^t(\phi_u)\right] \tag{4a}$$

$$\text{s.t.} \quad 1 \le \psi \le N_{\max}, \tag{4b}$$

$$0 \le N \le l_u^{\max}, \tag{4c}$$

$$E_0 \le E_u^t. \tag{4d}$$

We are particularly focused on achieving the highest possible success rate. However, we do not engage in separate power optimization or channel conditions. Instead, we maintain a constant transmit power setting that applies uniformly across all channel realizations. Additionally, we do not prescribe how users should utilize the energy they harvest; our primary concern is ensuring that each user receives sufficient energy to transmit a specified number of packet replicas. The exact number of replicas is contingent upon the energy available in each user, with a minimum requirement of at least one packet transmission per user. If a user has surplus energy, it has the flexibility to transmit additional replicas. Furthermore, we do not specify the precise number of replicas each user should send in each time frame but rather establish a maximum limit on the number of replicas that can be transmitted. Moreover, the allocation of packets to data slots is not predetermined; instead, it is determined randomly, avoiding fixed patterns.

The problem present in (4a) can be approached as a single-agent scenario, wherein a learning emergent protocol comes into consideration. In this setup, each individual agent strives to learn the most suitable method for transmitting replicas, all while taking their energy limitations and the quantity of sent packets into careful consideration. To tackle this problem effectively, we can employ the techniques of RL as *Q*-learning.

IV. RL policy for packet replication

In reinforcement learning (RL), a learning agent interacts with an environment to solve a sequential decision-making problem model as a discrete time markov decision process (MDP). Formally, MDP is defined as a tuple (S, A, P, R). Here, S is the set of all possible states, A is the set of actions, P is the transition probability function, and R is a reward function. In the EH-oriented data transmission scheme each user at time frame t is an agent with state $s_u^t \in S_u$. We outline our MDP as

1) State Space S_u: A state is defined as the energy level of each user. At time frame t, state $s_u^t \in S_u$ corresponds to the amount of energy at user u. Thus, we have a discrete state space: $S = [0, \xi_u, 2\xi_u, \ldots, \omega\xi_u]$.
2) Action Space A_u: An action is defined as the number of replicas sent by each user. For example, action means user u sends a_u^t replicas at time frame t. Furthermore, each user is able to send at most one packet per time frame. Consequently, the maximum number of packets each user sends is $N_u + 1$, and the maximum number

of replicas each user sends is $l_u^{\max}$. Therefore, we have a discrete action space: $A_u = [0, 1, \ldots, N_u]$.

3) Transition Probability P: The transition probability between states is unknown.
4) Reward R_u: The reward is defined as the number of successfully decoded packets using IRSA channel protocol in PB at each time frame as the reward.

Algorithm 1 Pseudocode for Q-learning

```
Initialize μ, δ and Q(s_u, a_u) randomly
for each frame t ∈ τ do
  for user u ∈ U do
    Observe the s_u^t and generate x ∼ U[0,1]
    if x < ε then
      Select an action randomly
    else
      Select an action: a_u^t(s_u^t) = arg max_{a∈A} Q(s_u^t, a)
    end if
    Randomly select data slots for action a_u^t
    Decode the packets using IRSA
    Collect the reward R_u^t and send to respective user
    Update the Q-value Q(s_u^t, a_u^t) as show in (5)
  end for
end for
```

A. Q-Learning

To solve the proposed MDP problem we employ a RL algorithm known as Q-learning. It is a model-free algorithm, meaning it does not require prior knowledge of the underlying system dynamics. Q-learning method determines the optimal policy that maximizes a given reward. Q-learning uses a table, known as the Q-table, to store action-value estimates for each state-action pair in the MDP [17]. The action-value estimate, denoted by $Q(s_u^t, a_u^t)$, represents the expected cumulative reward that an agent will receive by taking action a_u^t in state s_u^t and following a specific policy.

The Q-learning algorithm relies on iteratively updating the Q-values based on the observed rewards and the agent's experiences. At each time step, the agent selects an action based on an exploration-exploitation strategy, such as ϵ-greedy, which balances between trying new actions and exploiting the current best-known actions. After taking an action, the agent receives a reward and transitions to a new state. The Q-value for the previous state-action pair is then updated using Bellman's equation as follows [17]:

$$Q(s_u^t, a_u^t) = (1-\mu)Q(s_u^t, a_u^t) + \mu(R_u^t + \delta \max_{a \in A}(Q(s_u^{t+1}, a_u))), \tag{5}$$

where μ is the learning rate factor and δ is discount factor.

B. Learning Algorithm

We implement an independent Q-learning algorithm where the users exchange information with the BS, and each user employs Q-learning to learn its own policy. We adopt ϵ-greedy for action selection [17]. Where with probability $(1-\epsilon)$ the agent will select the highest Q-value action. Otherwise, the agent will randomly select an action. To ensure convergence, we decay the ϵ value over time. Let τ be the total number of time frames. Algorithm 1 presents a standard Q-learning algorithm for single agent RL. First, it initializes the learning parameters and $Q(s_u^t, a_u^t)$ randomly. Then, for each frame, t each user u observes the current energy state s_u^t and generates a random number $x \in [0, 1]$. Then to select an action we use ϵ-greedy, where with probability ϵ, where each user selects randomly an action. Otherwise, the user selects the action with the highest Q-value. Based on the previously selected action a_u^t, each user randomly selects data slots for the action. Then, IRSA is applied to decode the packets and BS collects the reward R_u^t. Then, user u observes its next state s_u^{t+1} and receives the reward from BS to find the highest Q-value for the new state. Finally, the users update the Q-value.

Table I: Simulation Parameters

Description	Value
EH saturation level	$\mathcal{W} = 10.73$ mW
EH unitless constants	$c_0 = 0.2308$ $c_1 = 5.365$
Distance between PB and user	$d = 3$ m
Distance between BS and user	$H = 70$ m
Number of antennas at the PB	$M \in \{4, 8\}$ antennas
Number of users	$u = 4$
LOS parameters	$\kappa = 2$ dB
Path-loss exponents	$\alpha = 2.7$
Carrier frequencies	$f = 2.5$ GHz
Charging slot length	$t_{\mathrm{C}} = 1$ ms
Data slot length	$t_{\mathrm{D}} = 1$ ms
Maximum number of packets	$N = 5$
Packet size	$L = 21$ bytes
Data Transmission Power	$P = 10$ mW
Transmit Power of the PB	$P_{\mathrm{b}} = 1$ W
Learning rate factor	$\mu = 0.1$
Discount factor	$\delta = 0.1$
ϵ-greedy value	$\epsilon = 0.5$

V. Numerical Results

In this section, we present numerical results to validate our scheme. We employ default simulation parameters, as enumerated in Table I unless specified otherwise. The propose RL-based data transmission scheme that uses an IRSA protocol is compared with a baseline that uses an CRDSA data transmission scheme without learning [10]. This baseline operates on the premise that the user always sends two replicas. However, if the user does not accumulate sufficient energy to transmit both replicas, only the main packet will be sent. We also compare the A-CSI scheme, based on average channel estimation, and the F-CSI scheme, which assumes the availability of perfect instantaneous channel information for the user-PB link. It is important to mention that all the results presented here are based on averaging data gathered from ten simulation runs, with each run spanning 5000 time frames. In Fig. 3, we observe the convergence behavior of the Q-learning. We conduct 1000 iterations, each comprising 450 time frames. Notably, it becomes evident that a system

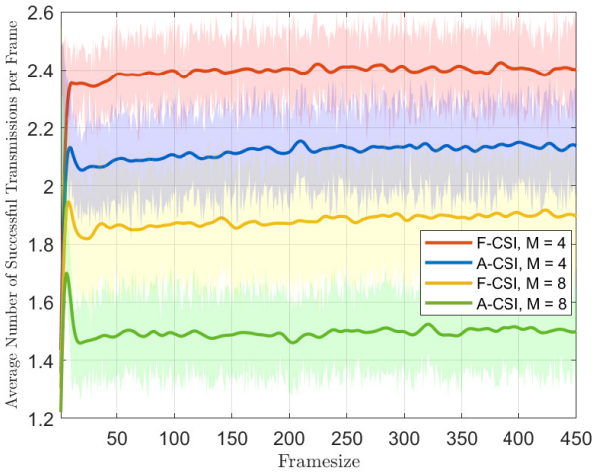

Figure 3: Convergence curve for F-CSI and A-CSI.

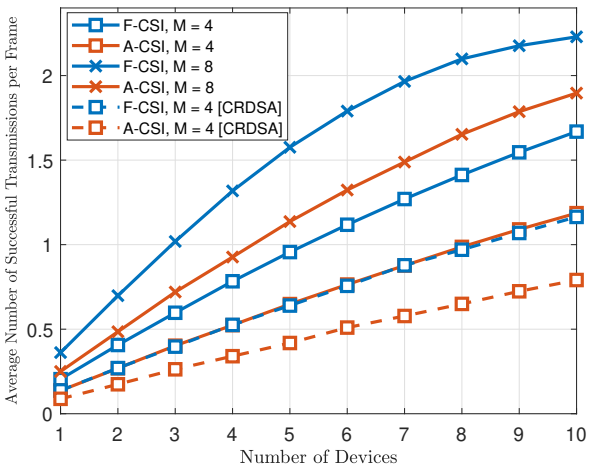

Figure 5: Impact of number of users on the number of successful transmissions per frame.

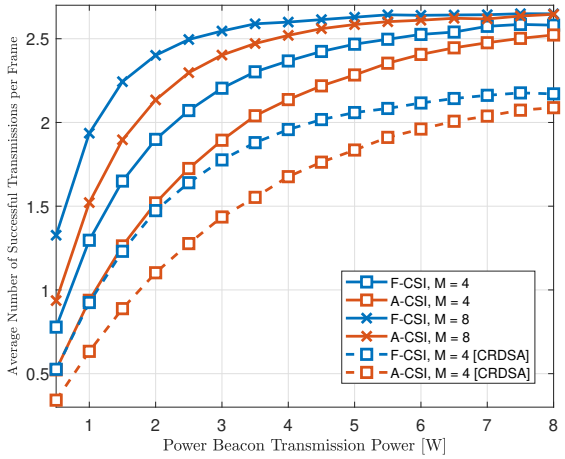

Figure 4: Impact of different PB transmit power on the number of successful transmissions per frame.

with a greater number of antennas tends to converge quickly if compared with the same scheme with fewer antennas.

In Fig. 4, we explore the relationship between the transmission power of the PB signal and the average number of successful transmissions per frame. It is evident that the performance of all schemes experiences improvement as the PB transmission power is increased. As observed, we can affirm that the system with the learning will give 18% more successful transmission if compared with the same scheme without the learning. Specifically, both F-CSI and A-CSI schemes with $M = 8$ saturate at an average of 2.6 successful transmissions per frame when the transmission power reaches 6 W. Additionally, an interesting observation emerges as we scrutinize the gap between the A-CSI and F-CSI schemes. This gap gradually diminishes as the PB transmission power increases. For instance, when $P_{\mathrm{b}} = 2$ W, the F-CSI scheme achieves 19% more successful packet transmissions compared to the A-CSI scheme. However, this advantage decreases, and when $P_{\mathrm{b}} = 4$ W, the F-CSI scheme outperforms the A-CSI scheme by 9% in terms of successful transmissions.

Figure 5 provides insight into how the number of users sharing information in the system impacts the number of successful transmissions per frame when the transmission power by PB is $P_{\mathrm{b}} = 27$ dB. We observe an increase in the number of successful transmissions as the number of devices increases. Notably, when we compare A-CSI with Q-learning and a parameter $M = 4$, its performance closely matches that of F-CSI CRDSA with $M = 8$. This observation suggests that by using a simpler hardware structure and leveraging Q-learning for average state estimation, it is possible to achieve an equivalent level of successful packet transmission.

Next, in Fig. 6, we can see how the duration of charging time affects the number of successful transmissions per frame. It is important to note that the performance of all the different schemes improves as the charging time increases. Upon closer observation, we can identify a critical charging time that leads to the highest number of successful transmissions for each scheme. When comparing this system with the lack of a learning component, it achieves a notable 17% increase in successful transmissions per frame compared to an equivalent system that utilizes the CRDSA channel protocol. For instance, in the case of F-CSI and A-CSI schemes with $M = 8$, we achieve a total of 2.66 successful transmissions when the charging time exceeds 4 ms. This observation strongly supports the idea that there is a minimum required charging time to attain the optimal number of successful transmissions per frame.

Finally, in Fig. 7, we depict the LOS factor's variation in relation to the average number of successful transmissions per frame. Notably, as we examine the graph, we observe a significant reduction in the gap between both schemes as the parameter κ increases. In a scenario with severe fading conditions, such as when $\kappa = 1$ dB, we find that the F-CSI

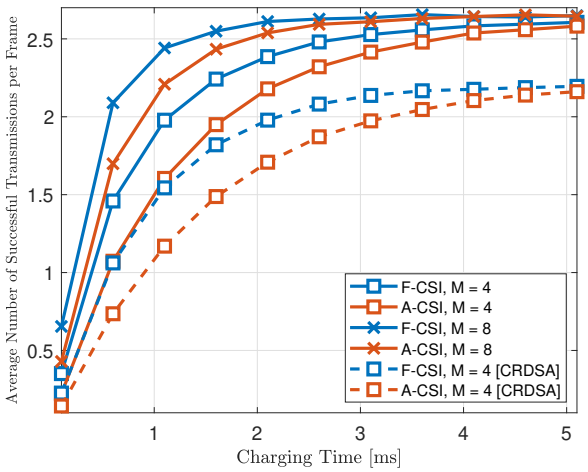

Figure 6: Impact of charging time on the number of successful transmissions per frame.

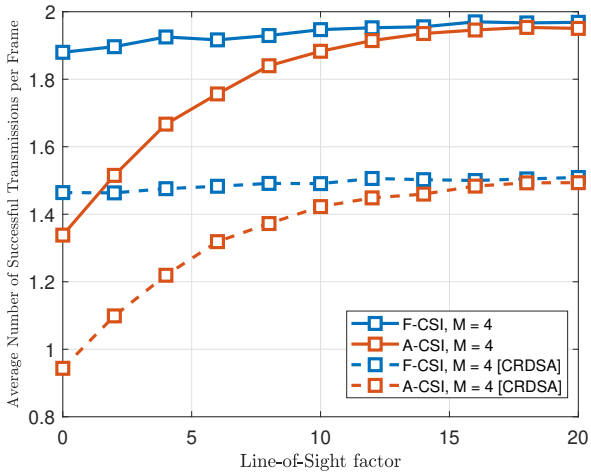

Figure 7: Impact of LOS factor on the number of successful transmissions per frame.

scheme with $M = 4$ antennas achieves a 25% higher number of successful transmissions compared to the A-CSI scheme with an equivalent number of antennas. This performance gap diminishes to 9% when $\kappa = 6$ dB. Additionally, it is worth noting that the F-CSI scheme, without the inclusion of action learning, outperforms a A-CSI system with learning only when κ is less than 2 dB. These results underscore the efficacy of A-CSI as a viable beamforming option, particularly when the PB can be strategically positioned in a favorable configuration, enjoying a strong LOS connection with the user nodes.

VI. Conclusion

This paper presents a EH-oriented data transmission scheme where a set of single-antenna users is charged by their own PB using A-CSI beamforming. After the EH, each user transmits the data to a BS using the IRSA protocol. First, we characterize the closed-form expression for the energy model. The distributed Q-learning algorithm finds the optimal policy that maximizes the number of successful transmissions per frame under energy constraints. As the numerical results show the Q-learning method increases the number of successful transmissions per frame in 18% if compared with the same scheme using CRDSA channel protocol without the learning process. Also, we can affirm that A-CSI beamforming in the worst case achieved 19% less successful transmission than the F-CSI. The optimal policy given by Q-learning solution ensures a higher number of successful transmissions per frame for each user, than in the case of a transmission scheme not optimized with an RL policy.

References

[1] Cisco, "Cisco annual internet report, 2018 - 2023," 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] D. C. Nguyen *et al.*, "6G Internet of Things: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 359–383, 2022.

[3] K. W. Choi *et al.*, "Distributed wireless power transfer system for Internet of Things devices," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2657–2671, 2018.

[4] A.-N. Nguyen *et al.*, "Performance analysis and optimization for IoT mobile edge computing networks with rf energy harvesting and UAV relaying," *IEEE Access*, vol. 10, pp. 21 526–21 540, 2022.

[5] T.-H. Vu *et al.*, "Performance evaluation of power-beacon-assisted wireless-powered NOMA IoT-based systems," *IEEE Internet of Things Journal*, 2021.

[6] Z. Chen *et al.*, "Energy efficiency optimization for irregular repetition slotted ALOHA-based massive access," *IEEE Wireless Communications Letters*, vol. 11, no. 5, pp. 982–986, 2022.

[7] U. Demirhan *et al.*, "Irregular repetition slotted ALOHA with energy harvesting nodes," *IEEE Transactions on Wireless Communications*, vol. 18, no. 9, pp. 4505–4517, 2019.

[8] C. d. V. Silva *et al.*, "Slotted ALOHA for wireless powered resource-constrained networks," in *IEEE International Conference on Communications*, 2021, pp. 1–6.

[9] Y. Li *et al.*, "Energy-aware irregular slotted aloha methods for wireless-powered IoT networks," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 11 784–11 795, 2022.

[10] A. Meloni *et al.*, "CRDSA, CRDSA++ and IRSA: Stability and performance evaluation," in *2012 6th Advanced Satellite Multimedia Systems Conference and 12th Signal Processing for Space Communications Workshop*, 2012, pp. 220–225.

[11] M. Ghanbarinejad *et al.*, "Irregular repetition slotted ALOHA with multiuser detection," in *2013 10th Annual Conference on Wireless On-demand Network Systems and Services*, 2013, pp. 201–205.

[12] A. Goldsmith, *Wireless Communications*, 1st ed. Cambridge University Press, 2005.

[13] O. A. López *et al.*, "A low-complexity beamforming design for multiuser wireless energy transfer," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 58–62, 2021.

[14] T. Lo, "Maximum ratio transmission," *IEEE Transactions on Communications*, vol. 47, no. 10, pp. 1458–1461, 1999.

[15] E. Boshkovska *et al.*, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2082–2085, 2015.

[16] B. Clerckx *et al.*, "Fundamentals of wireless information and power transfer: From RF energy harvester models to signal and system designs," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 1, pp. 4–33, 2019.

[17] R. S. Sutton *et al.*, "Reinforcement learning," *Journal of Cognitive Neuroscience*, vol. 11, no. 1, pp. 126–134, 1999.