

Auction-Based RIS Allocation With DRL: Controlling the Cost-Performance Trade-Off

1st Martin Mark Zan
Institute of Telecommunications
TU Wien
Vienna, Austria
martin.zan@tuwien.ac.at

2nd Stefan Schwarz
Institute of Telecommunications
TU Wien
Vienna, Austria
stefan.schwarz@tuwien.ac.at

Abstract—We study the allocation of reconfigurable intelligent surfaces (RISs) in a multi-cell wireless network, where base stations compete for control of shared RIS units deployed at the cell edges. These RISs, provided by an independent operator, are dynamically leased to the highest bidder using a simultaneously ascending auction format. Each base station estimates the utility of acquiring additional RISs based on macroscopic channel parameters, enabling a scalable and low-overhead allocation mechanism. To optimize the bidding behavior, we integrate deep reinforcement learning (DRL) agents that learn to maximize performance while adhering to budget constraints. Through simulations in clustered cell-edge environments, we demonstrate that reinforcement learning (RL)-based bidding significantly outperforms heuristic strategies, achieving optimal trade-offs between cost and spectral efficiency. Furthermore, we introduce a tunable parameter that governs the bidding aggressiveness of RL agents, enabling a flexible control of the trade-off between network performance and expenditure. Our results highlight the potential of combining auction-based allocation with adaptive RL mechanisms for efficient and fair utilization of RISs in next-generation wireless networks.

Index Terms—Reconfigurable Intelligent Surfaces, Resource Allocation, Auctions, Reinforcement Learning, Multi-Agent Systems, 6G.

I. INTRODUCTION

IN the pursuit of beyond-5G and 6G wireless networks, improving spectral and energy efficiency has become a central research objective. Reconfigurable intelligent surfaces (RIS) have recently emerged as a promising technology to reshape the radio environment and enhance coverage in challenging propagation conditions [1]. By enabling programmable reflection of incident signals, RISs provide a cost- and energy-efficient means to strengthen desired links and suppress interference, thereby complementing conventional base station and user equipment capabilities [2], [3]. Their practical use, however, raises fundamental design questions, including how to deploy RISs in realistic network topologies and how to efficiently allocate them to different transmitters and users.

In this work, we concentrate on scenarios where multiple base stations compete for control of RISs located near the cell boundary. Such situations naturally arise at the cell edge,

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT4490824. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

where the positioning of RISs can have a decisive impact on system performance. This setting introduces a resource allocation problem: when several transmitters contend for the same RISs, an effective mechanism is required to determine access in a fair and efficient manner. To address this challenge, we consider a market-inspired approach in which RISs are not permanently assigned but are dynamically leased to the highest bidder through an ascending auction format. This mechanism provides a low-complexity and scalable alternative to combinatorial allocation methods [4], while capturing the competitive interactions between base stations. Similar auction-based approaches have been successfully applied in spectrum allocation, where simultaneous ascending auctions enable efficient and competitive distribution of scarce frequency resources [5], [6]. A related approach was recently investigated in [7], where an auction-based RIS allocation was studied in a multi-operator setting. Building on this framework, we further employ reinforcement learning (RL) to optimize the bidding strategies of individual agents. By doing so, the agents can adapt their behavior to the evolving auction environment, selectively targeting high-value RISs and avoiding inefficient bidding. Our study demonstrates that this combination of auction-based allocation and RL-driven bidding achieves superior trade-offs between cost and performance compared to heuristic approaches.

Notation: The complex Gaussian distribution with mean μ and covariance matrix C is denoted as $\mathcal{CN}(\mu, C)$. The uniform distribution over the interval $[a, b]$ is written as $\mathcal{U}(a, b)$. The transpose and Hermitian transpose of a vector \mathbf{x} are \mathbf{x}^T and \mathbf{x}^H , respectively, and its i -th element is $\mathbf{x}[i]$. The Euclidean norm of \mathbf{x} is $\|\mathbf{x}\|$. The size of a set \mathcal{X} is $|\mathcal{X}|$, and the empty set is \emptyset . The expectation of a random variable r is $\mathbb{E}[r]$, and the phase of a complex number z is $\arg(z)$.

II. SYSTEM MODEL

Let us have N_{BS} base stations which serve N_{UE} users in parallel over N_{RIS} RISs. We assume that each base station is equipped with M_{BS} antennas and users are equipped with a single antenna, while each RIS consists of M_{RIS} reconfigurable elements.

We have a direct channel between each base station b and each user u , denoted as $\mathbf{h}_{u,b}^{\text{direct}} \in \mathbb{C}^{M_{\text{BS}} \times 1}$. This channel is written as

$$\mathbf{h}_{u,b}^{\text{direct}} = \gamma_{u,b} \mathbf{g}_{u,b}. \quad (1)$$

The channel gain $\gamma_{u,b}$ depends on the distance between the user and the base station. The fading component $\mathbf{g}_{u,b}$ follows a zero-mean unit-variance complex Gaussian distribution, and $\mathbb{E}[\|\mathbf{g}_{u,b}\|^2] = M_{\text{BS}}$. The normalization condition on the channel is $\mathbb{E}[\|\mathbf{h}_{u,b}^{\text{direct}}\|^2] = \gamma_{u,b}^2 M_{\text{BS}}$. We assume that the base station to user link is non-line-of-sight (NLOS) and it is strongly shadowed, so that the line-of-sight (LOS) component can be neglected. This makes sense in this setup, since we are interested in RIS-aided transmissions.

In addition to this direct channel, we have RIS-assisted channels. These are composed of the channel between each base station and the RISs, the RISs' responses, and the channel between the RISs and the users.

Let us consider the channel between the base station b and RIS r : $\mathbf{H}_{r,b} \in \mathbb{C}^{M_{\text{RIS}} \times M_{\text{BS}}}$. This is a matrix valued channel, containing all channels between M_{BS} base station antennas and the M_{RIS} RIS elements. It is written as:

$$\mathbf{H}_{r,b} = \gamma_{r,b} \mathbf{a}(\psi_{r,b}) \mathbf{a}(\theta_{r,b})^T. \quad (2)$$

Here, $\gamma_{r,b}$ is the path gain between the base station and the RIS, $\mathbf{a}(\psi_{r,b})$ denotes the directional RIS response vector, and $\mathbf{a}(\theta_{r,b})$ is the directional base station response vector [8]. They depend on the angle-of-arrival $\psi_{r,b}$ of the signal at the RIS and the angle-of-departure $\theta_{r,b}$ of the signal at the base station. For simplicity, it is assumed that the base station to RIS channel is under strong LOS, so that the fading NLOS component is neglected. This assumption is reasonable, since RISs are typically deployed in locations that ensure unobstructed visibility to nearby base stations, for example, mounted on building facades or rooftops in urban environments.

The channel between RIS r and user u is $\mathbf{h}_{u,r} \in \mathbb{C}^{M_{\text{RIS}} \times 1}$. We use a Rician channel model, in which the channel is composed of a non-fading LOS component and a random Gaussian fading NLOS component. The relative strength of the two channel components is characterized by the Rician K -factor, which is higher under LOS conditions and lower under NLOS conditions. It is written as follows

$$\mathbf{h}_{u,r} = \gamma_{u,r} \left(\sqrt{\frac{K_{u,r}}{1+K_{u,r}}} \mathbf{a}(\theta_{u,r}) + \sqrt{\frac{1}{1+K_{u,r}}} \mathbf{g}_{u,r} \right). \quad (3)$$

The channel gain $\gamma_{u,r}$ depends on the line-of-sight conditions of the link (higher gain under line-of-sight, lower under non-line-of-sight) and on the distance between the user and the RIS. The non-fading component $\mathbf{a}(\theta_{u,r})$ is the directional base station response vector [8], and $\mathbb{E}[\|\mathbf{a}(\theta_{u,r})\|^2] = M_{\text{RIS}}$. The angle-of-departure of the signal at the RIS is $\theta_{u,r}$. Let us define $k_{u,r} = \sqrt{\frac{K_{u,r}}{1+K_{u,r}}}$ and $\bar{k}_{u,r} = \sqrt{\frac{1}{1+K_{u,r}}}$. The vector $\mathbf{g}_{u,r}$ is defined analogously as in the direct channel.

The RIS infrastructure is managed by an independent provider who leases the units on-demand to the highest bidder. Once a base station gains control of a RIS, it can configure

its phase-response to enhance signal quality for its associated users. We consider a diagonal RIS, so that the RIS response can be characterized by a diagonal phase-shift matrix $\Phi_r = \text{diag}(e^{j\phi_{r,1}}, \dots, e^{j\phi_{r,M_{\text{RIS}}}}) \in \mathbb{C}^{M_{\text{RIS}} \times M_{\text{RIS}}}$. With this, the overall RIS-assisted channel with all RISs r is $\mathbf{h}_{u,b}^{\text{indirect}} \in \mathbb{C}^{M_{\text{BS}} \times 1}$, which is defined as:

$$\mathbf{h}_{u,b}^{\text{indirect}} = \sum_{r=1}^{N_{\text{RIS}}} (\mathbf{h}_{u,r}^T \Phi_r \mathbf{H}_{r,b})^T. \quad (4)$$

For RIS r with elements indexed by i , the optimized phase-shift is given by

$$\phi_{r,i} = -(\arg(\mathbf{a}(\theta_{u,r})[i]) + \arg(\mathbf{a}(\theta_{r,d})[i])). \quad (5)$$

We assume that random scattering components $\mathbf{g}_{u,r}$ change so quickly that we cannot estimate them reliably. Therefore, they cannot be phase-aligned, and they contribute incoherently. If a RIS is not assigned to the serving base station, its elements are modeled as random phases, i.e., $\phi_{r,i} \sim \mathcal{U}(0, 2\pi)$.

Furthermore, the total direct + assisted channel $\mathbf{h}_{u,b} \in \mathbb{C}^{M_{\text{BS}} \times 1}$ is

$$\mathbf{h}_{u,b} = \mathbf{h}_{u,b}^{\text{direct}} + \mathbf{h}_{u,b}^{\text{indirect}}. \quad (6)$$

We consider a system model in which each base station serves multiple users using time-orthogonal resources, such that only one user is scheduled per time slot. Based on these assumptions, the frequency-flat single-input single-output (SISO) downlink input-output relation for user u can be expressed as follows:

$$y_u = \mathbf{h}_{u,d}^T \mathbf{f}_{u,d} x_u + \sum_{b \neq d} \mathbf{h}_{u,b}^T \mathbf{f}_{j_b,b} x_{j_b} + n_u, \quad (7)$$

where d denotes the index of the base station serving user u , $\mathbf{f}_{u,d} \in \mathbb{C}^{M_{\text{BS}} \times 1}$ is the beamforming vector applied to transmit to user u at base station d . $P_{u,b}$ is the power used to transmit to user u from base station b and is considered as part of the beamformer, such that $\mathbb{E}[\|\mathbf{f}_{u,b}\|^2] = P_{u,b}$. x_u is the symbol intended for user u ; normalization is applied to transmit symbols, i.e. $\mathbb{E}[|x_u|^2] = 1$. j_b denotes the index of the user served by base station b and $n_u \sim \mathcal{CN}(0, \sigma_n^2)$ is noise. We also assume that the beamforming is based only on the directional parts of the channel, the angle-dependent array steering vectors, since the NLOS part changes too quickly to be reliably estimated.

The users' instantaneous signal to interference and noise ratios (SINRs) $\beta_u^{(d)}$ and achievable rates $r_u^{(d)}$ are:

$$\beta_u^{(d)} = \frac{|\mathbf{h}_{u,d}^T \mathbf{f}_{u,d}|^2}{\sigma_n^2 + \sum_{b \neq d} |\mathbf{h}_{u,b}^T \mathbf{f}_{j_b,b}|^2}, \quad r_u^{(d)} = \log_2(1 + \beta_u^{(d)}). \quad (8)$$

III. UTILITY ESTIMATION

In order to allocate RISs, each base station needs to evaluate the potential performance gain that can be achieved when certain RISs are assigned to it. Since perfect channel state information is not available before the RISs are configured, we rely on macroscopic channel properties to obtain coarse estimates of the achievable SINR and sum rates. These estimates are then used to define a utility function which guides the allocation.

A. SINR Estimation

Based on the channel model in Section II, we estimate the users' SINR using the macroscopic channel parameters, by approximating the instantaneous power values in (8) with their expected values. For sufficiently large antenna arrays and RISs, this approximation provides a good estimate due to the law of large numbers. We decompose the received powers into direct, coherent RIS-assisted, and incoherent RIS-assisted components, as well as direct and RIS-assisted interference. The estimated SINR for user u served by base station d is

$$\hat{\beta}_u^{(d)} = \frac{p_d + p_c + p_i}{\sigma_n^2 + i_d + i_i}, \quad (9)$$

where

$$\begin{aligned} p_d &= \gamma_{u,d}^2 P_{u,d}, \\ p_c &= \left(\sum_{r \in R^{(d)}} \gamma_{u,r} \gamma_{r,d} k_{u,r} \sqrt{\frac{P_{u,d} M_{BS}}{|R^{(d)}|}} M_{RIS} \right)^2, \\ p_i &= \sum_{r \in R^{(d)}} \gamma_{u,r}^2 \gamma_{r,d}^2 k_{u,r}^2 \frac{P_{u,d} M_{BS}}{|R^{(d)}|} M_{RIS}, \\ i_d &= \sum_{b \neq d} \gamma_{u,b}^2 P_{j_b,b}, \\ i_i &= \sum_{b \neq d} \sum_{r \notin R^{(d)}} \gamma_{u,r}^2 \gamma_{b,r}^2 P_{j_b,b} M_{RIS}. \end{aligned}$$

Here $R^{(d)}$ denotes the set of RISs assigned to base station d . The numerator consists of the direct signal p_d , the coherent RIS-assisted components p_c , and the incoherent RIS-assisted components p_i . The denominator includes noise σ_n^2 , direct interference i_d from non-serving base stations $b \neq d$, and RIS-assisted interference i_i caused by RISs not assigned to the serving base station d .

We assume that the beamformer of each base station points towards its assigned RISs with uniform power distribution. If no RIS is assigned, the beamformer is chosen to be isotropic. In this case, we only have a channel with K -factor equal to 0. Since this random Rayleigh fading channel is not known to the base station, we cannot match the beamformer to this channel and therefore can only select it randomly. Interfering base stations are also assumed to employ isotropic beamforming, since their RIS assignments are unknown to base station d . Finally, we assume that the directional base station response vectors are asymptotically orthogonal. The estimated achievable rate then follows as

$$\hat{r}_u^{(d)} = \log_2 \left(1 + \hat{\beta}_u^{(d)} \right). \quad (10)$$

B. Utility Function

We adopt a utility function that measures the percentage sum-rate improvement compared to the case without RISs. The utility for a given allocation $R^{(b)}$ is defined as

$$U^{(b)}(R^{(b)}) = \frac{\sum_{u=1}^{N_{UE}^{(b)}} \hat{r}_u^{(b)}(R^{(b)})}{\sum_{u=1}^{N_{UE}^{(b)}} \hat{r}_u^{(b)}(\emptyset)} - 1, \quad (11)$$

where $\hat{r}_u^{(b)}(R^{(b)})$ denotes the estimated rate of user u when RISs $R^{(b)}$ are allocated [7].

C. Auction Format

The allocation of RISs to base stations is performed using a simultaneously ascending ("Japanese" forward) auction format, as introduced in [7]. This mechanism was originally motivated by the need for a low-complexity alternative to combinatorial auctions such as the Vickrey-Clarke-Groves (VCG) mechanism [4]. In each round t , the auctioneer increases the price p_t of each RIS by a fixed increment $\Delta_p = p_t - p_{t-1}$. Base stations then indicate their willingness to pay the current price for individual RIS units using a binary "bid" vector $b_t^{(b)} \in \{0, 1\}^{N_{RIS}}$. If a RIS receives only one bid, it is allocated to the corresponding base station; if a RIS receives more than one bid, the bidding continues in the next round; if the RIS receives zero bids then it remains unassigned, and its phase-response is set randomly. To ensure consistency, an activity rule is enforced, i.e., a base station cannot rejoin the bidding for a RIS if it did not bid for it in the previous round [9]. We adopt this mechanism in our setup without modification.

IV. BIDDING STRATEGIES

At each auction round t , base stations must decide which RISs are worth bidding for, based on the current price p_t and their estimated utility. Let $R_{t-1}^{(b)}$ denote the set of RISs already allocated in previous rounds to base station b . The remaining unassigned RISs form the set $R_t = \{1, \dots, N_{RIS}\} \setminus \bigcup_b R_{t-1}^{(b)}$. Each base station b then evaluates the potential gain of adding an individual RIS $r \in R_t$ to its current allocation $R_{t-1}^{(b)}$.

Ideally, this decision would be based on the expected utility of all possible combinations of the remaining RISs. However, such a combinatorial evaluation becomes infeasible when the number of RISs is large, due to the exponential number of subsets. To keep the computation tractable, we adopt a simplified approach, where each base station evaluates the benefit of acquiring a single additional RIS r , assuming no other RIS will be acquired in the same round. The estimated marginal value of RIS r is then given by:

$$V_t^{(b)}(r) = U^{(b)}(R_{t-1}^{(b)} \cup \{r\}) - U^{(b)}(R_{t-1}^{(b)}). \quad (12)$$

To standardize the scale of utility gains across different bidding scenarios, the estimated values are normalized as:

$$V_t^{(b)}(r) \leftarrow \frac{V_t^{(b)}(r)}{\max_{r'} V_t^{(b)}(r')}, \quad (13)$$

resulting in values bounded within the interval $[-1, 1]$. A positive value indicates a potential improvement in performance, while a negative value suggests a possible degradation. Negative values can arise due to our assumption of uniform power allocation over RIS-directed transmit beams at the base station.

Based on these estimated values, different bidding strategies can be applied. In the following subsections, we consider three approaches: two heuristic strategies based on dominant

bidding behavior, and a reinforcement learning based strategy that adapts over time to maximize long-term reward.

A. Heuristic Bidding

As a baseline approach, we implement a simple greedy bidding strategy. At each auction round t , base station b computes the value $V_t^{(b)}(r)$ for each available RIS $r \in R_t$, based on the definition in (13). After normalization, the RISs are ranked in descending order of estimated value.

To determine how many RISs to bid for, we calculate the remaining budget $B_t^{(b)}$ as the difference between the initial budget $B_0^{(b)}$ and the total cost already paid for previously assigned RISs $R_{t-1}^{(b)}$. A conservative bidding rule is then applied: the base station places bids for the $\lfloor B_t^{(b)}/p_t \rfloor$ most valuable RISs at round t . Although this strategy does not account for long-term optimization or strategic interaction across rounds, it provides a low-complexity and deterministic baseline that reflects greedy decision-making based solely on immediate value.

In addition to this value-based approach, we also evaluate a simpler distance-based heuristic. In this variant, the value function is replaced with a distance-derived metric defined as

$$VD_t^{(b)}(r) = \frac{1}{\text{dist}(b, r) + \varepsilon}, \quad (14)$$

where $\text{dist}(b, r)$ is the Euclidean distance between base station b and RIS r , and ε is a small constant to avoid invalid division. The rest of the bidding procedure remains unchanged. This strategy assumes that proximity correlates with utility, and serves as a lower-complexity baseline that does not rely on SINR estimation.

B. RL-Based Bidding

Alternatively to these heuristic strategies, we explore the use of RL to derive optimized bidding strategies. Each base station operates an independent RL agent, trained to make decisions based only on its local observations, without coordination or information sharing among other base stations. In the following, we define the structure of the environment used for training these agents, including the state and observation spaces, action representation, and the reward function.

a) States: The complete state of the auction environment at time step t is defined as:

$$S_t = \left(p_t, \left\{ V_t^{(b)}(r), B_t^{(b)} \mid \forall b, r \right\} \right). \quad (15)$$

To ensure a fixed-length state representation compatible with RL algorithms, we assign $V_t^{(b)}(r) = 0$ for all RISs r that:

- are no longer available (i.e., already assigned),
- were not bid on by base station b in the previous round (in line with the enforced activity rule),
- or for which the current price p_t exceeds the remaining budget of base station b .

This approach avoids explicitly tracking previously assigned RISs, while preserving the same semantics. It also enables scalability to environments with different numbers of RISs,

as unused RIS slots can be zero-padded. The environment maintains this full state internally, although each agent operates only on its local observation, reflecting the partially observable nature of the problem.

b) Observations: The observation available to base station b at time step t is given by:

$$O_t^{(b)} = \left(p_t, B_t^{(b)}, \left\{ V_t^{(b)}(r) \mid \forall r \right\} \right). \quad (16)$$

c) Actions: At each time step t , based on its observation $O_t^{(b)}$, each agent (i.e., base station b) outputs a bid vector $b_t^{(b)} \in \{0, 1\}^{N_{\text{RIS}}}$, where the binary entry $b_t^{(b)}[r] = 1$ indicates that the base station is willing to bid for RIS r at the current price p_t , and $b_t^{(b)}[r] = 0$ otherwise.

d) Rewards: The reward function is designed to guide each agent toward maximizing the value of its bids while staying within budget and avoiding wasteful spending. It consists of three components:

- **R1 (value of bids):** This term measures the total value of the RISs the agent bids on in a given round. It is computed as the sum of the estimated values $V_t^{(b)}(r)$ for all RISs r for which $b_t^{(b)}[r] = 1$:

$$R1 = \sum_{r=1}^{N_{\text{RIS}}} V_t^{(b)}(r) \cdot b_t^{(b)}[r]. \quad (17)$$

- **R2 (cost of bids):** This term penalizes agents for the total cost of their bids, scaled by the bid intensity parameter β . The cost is computed as the number of bids placed multiplied by the current price:

$$R2 = \beta \cdot p_t \cdot \sum_{r=1}^{N_{\text{RIS}}} b_t^{(b)}[r]. \quad (18)$$

- **R3 (overspending penalty):** This component introduces an additional penalty if the total bid cost exceeds the agent's remaining budget. The penalty is proportional to the overshoot and scaled by 2β :

$$R3 = 2\beta \cdot \max \left(p_t \cdot \sum_{r=1}^{N_{\text{RIS}}} b_t^{(b)}[r] - B_t^{(b)}, 0 \right). \quad (19)$$

The scaling factor of 2 in R3 is chosen to make overspending penalties clearly distinguishable from the regular cost term R2. This ensures that the agent perceives budget violations as more severe than ordinary expenditures, while avoiding excessively high penalties that would lead to overly conservative bidding behavior.

The final reward is then computed as:

$$r_t^{(b)} = R1 - R2 - R3 \quad (20)$$

This reward formulation encourages the agent to bid on high-value RISs (via R1), penalizes excessive or unaffordable bidding (via R2), and strictly discourages budget violations (via R3). Notably, the reward is computed before the actual auction outcome, allowing the agent to receive immediate feedback on its decisions and learn more efficiently. This leads

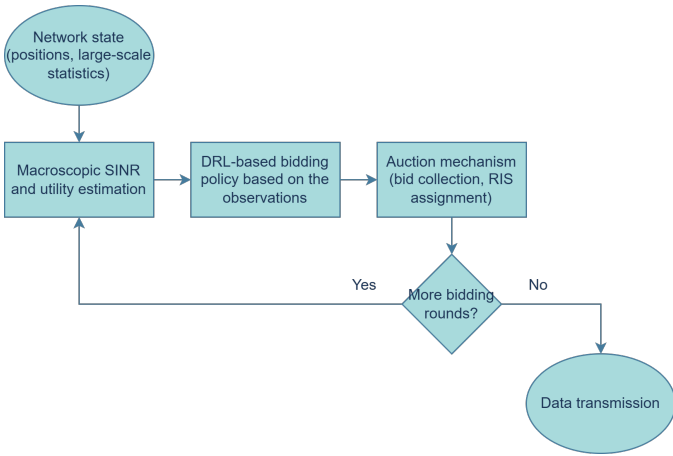


Fig. 1. System workflow of the proposed auction-based RIS allocation framework. Macroscopic SINR and utility estimates are used as inputs to a DRL-based bidding policy, which generates bids for the auction mechanism. The auction iterates over bidding rounds until a termination condition is met, after which data transmission is performed.

to a denser and more informative reward signal compared to outcome-based rewards. To enable generalization across environments with different budgets and prices, both the price p_t and the budget $B_t^{(b)}$ are normalized by the initial budget. The bid intensity parameter β is selected from a range of representative values to illustrate the cost-performance trade-off, and the values used in Section V are found to yield stable learning behavior and meaningful operating points.

We also evaluated several alternative reward structures. First, we tested using only the actual performance improvement (i.e., post-auction value gain), along with the cost and overspending penalty computed from the RISs that were actually won. However, the delayed feedback significantly hindered learning: agents were unable to reliably identify high-value RISs. Second, we experimented with nonlinear penalty functions in R3. Both quadratic and exponential penalties led to numerical instability: large bid costs produced disproportionately large negative rewards, causing the learning process to collapse.

e) Implementation: We implemented our RL-based auction environment using Gymnasium 1.1.1 [10] and PettingZoo 1.25.0 [11] for multi-agent support, along with SuperSuit 3.10.0 [12] for preprocessing and environment wrapping. Agents were trained using the Stable-Baselines3 2.6.0 [13] implementation of the Proximal Policy Optimization (PPO) algorithm [14], with default hyperparameters.

The environment supports a variable number of users, base stations, and RISs. Training is conducted episodically, where each episode corresponds to a full auction from start to finish. The episode length is variable and depends on the number of RISs and the bidding behavior of the agents. After each episode, a new environment is generated with randomized channel realizations and geometry. In practical scenarios with time-varying user locations, the trained bidding policy can be re-executed to quickly reassign RIS resources, without requiring retraining of the model. The system workflow can

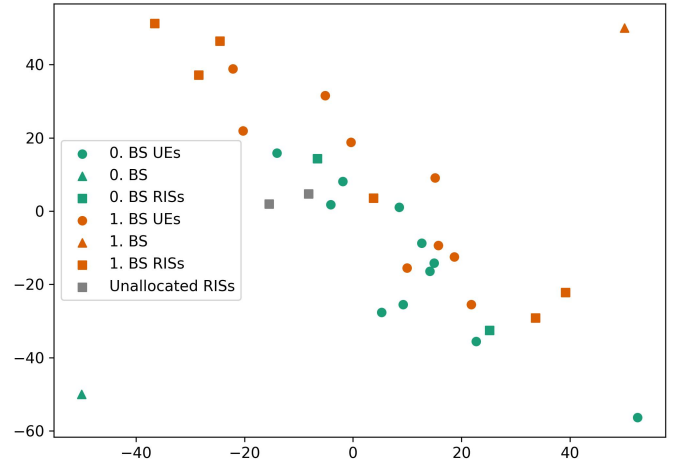


Fig. 2. Simulation geometry with two base stations (BS) at opposite ends, clustered users (UE), and RISs positioned along the cell boundary. The figure shows an example allocation of RISs obtained using a reinforcement learning agent with bid intensity parameter $\beta = 3$.

TABLE I
SIMULATION PARAMETERS

Carrier frequency	$f_c = 26$ GHz
Number of base stations	$N_{BS} = 2$
Number of base station antennas	$M_{BS} = 50$
Number of users	$N_{UE} = 20$
Number of RISs	$N_{RIS} = 10$
Number of RIS elements	$M_{RIS} = 250$
Transmit power per subcarrier	$P = 100$ mW
Subcarrier bandwidth	15 kHz
AWGN noise power spectral density	-174 dBm/Hz
Noise figure	6 dB
Path-loss exponent under LOS (NLOS)	2 (4.25)
K -factor under LOS (NLOS)	100 (3)
Distance-dependent LOS-probability	$p_{LOS}(d) = e^{-d/50}$
Shadow fading variance	10 dB
Auction initial price	$p_0 = 0.05$
Auction price increment	$\Delta p = 0.05$
Budget	$B_0^{(b)} = 1$

be seen on Fig. 1.

The implementation and the full list of dependencies and versions will be made available upon acceptance of the work at: <https://github.com/MartinMarkZan>.

V. SIMULATIONS

We evaluated the proposed auction-based allocation in a two-cell scenario focusing on the cell-edge region, where users experience the most ambiguous association conditions. We adopt a clustered geometry in which both RISs and users are concentrated close to the cell boundary. This reflects the practical observation that RIS-assisted links provide the strongest benefit when the RIS is positioned near either the transmitter or the receiver, rather than midway between them [15]. Specifically, we place $N_{BS} = 2$ base stations at opposite ends of a 100m^2 region of interest and generate correlated clusters of cell-edge users and RISs around the boundary line, as illustrated in Fig. 2 for an example RIS allocation obtained

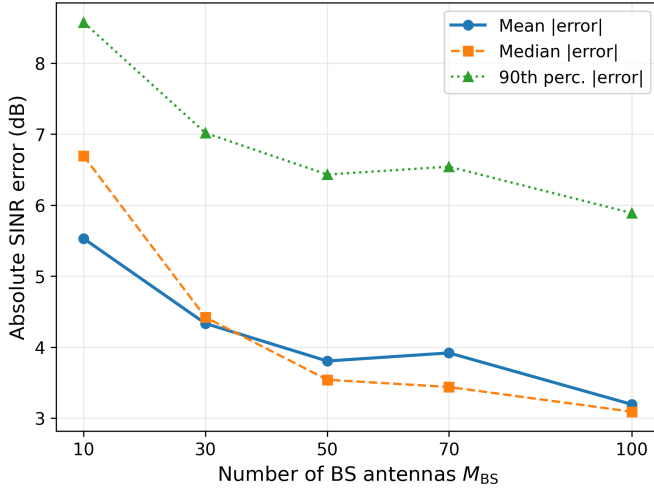


Fig. 3. Accuracy of the macroscopic SINR estimation as a function of the number of BS antennas M_{BS} . The figure reports the mean, median, and 90th percentile of the absolute error between the estimated and true SINR values.

with an RL agent at $\beta = 3$. The parameters of the simulations are listed in Table I.

The RL agent undergoes training through episodes. Training proceeds for up to $3 \cdot 10^6$ auction steps, with early stopping applied if no improvement in performance is observed over several evaluation intervals. During the evaluation, results are averaged over 200 macroscopic realizations (positions, path gains), each comprising 20 independent microscopic fading realizations (Rayleigh fading component).

To assess the accuracy of the proposed macroscopic SINR estimation, Fig. 3 reports multiple statistics of the absolute error between the estimated and true SINR values for different numbers of BS antennas M_{BS} for a single macroscopic realization. The results reveal a clear trend of decreasing estimation error as M_{BS} increases, indicating that the macroscopic approximation becomes more reliable for larger arrays. For smaller array sizes, a noticeable estimation mismatch persists.

To improve the credibility and reproducibility of the simulation results, we analyze the training dynamics of the DRL-based bidding agent. Fig. 4 shows the evolution of the reward value during learning. Following an initial transient phase with high variability, the reward improves steadily and converges to a stable plateau, indicating that the policy has converged prior to evaluation. A similar convergence behavior is observed for the other considered models, although larger β values lead to lower final reward levels.

Fig. 5 illustrates the trade-off between cost and achievable sum rate for the proposed RL-based allocation and the baseline strategies. Three reference points are considered: the two heuristic approaches introduced in Subsection IV-A, and the “Without RISs” case in which no RISs are allocated. From this figure, three main observations can be made. First, the absence of RISs leads to significantly lower performance, underlining the importance of RIS-assisted transmission. It should be noted that this effect is overemphasized in our simulations, due to

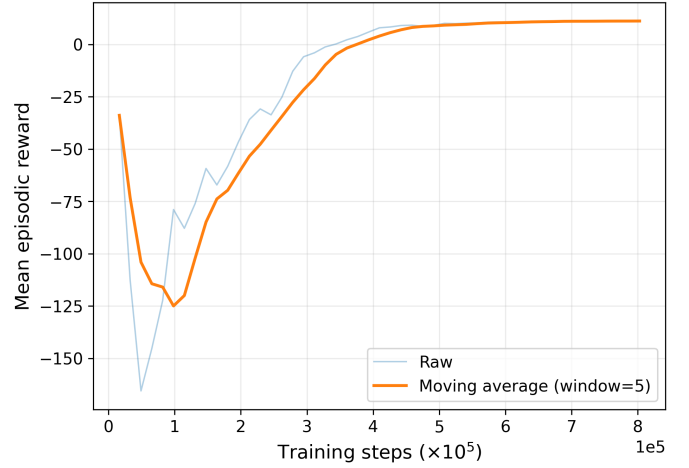


Fig. 4. Training convergence of the PPO-based bidding agent with $\beta = 2$. The figure shows the evolution of the training reward over environment interaction steps, with a moving-average smoothed curve (window size 5).

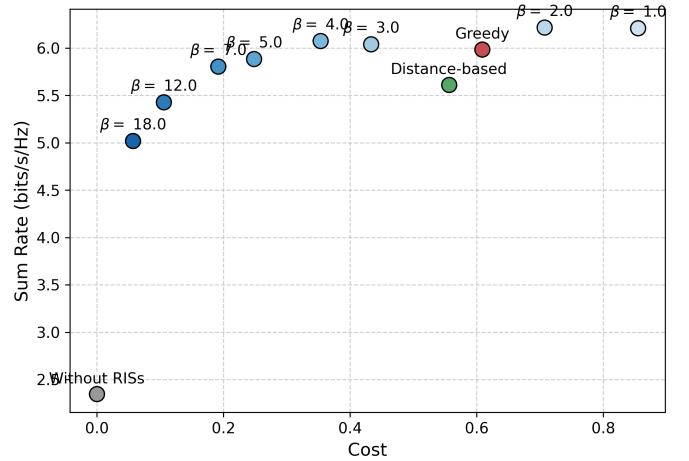


Fig. 5. Performance comparison of heuristics and RL-based models using different bid intensity (β) values. The figure illustrates the trade-off between cost and achievable rate.

the assumption that there is no direct LOS between users and base stations. Second, compared to the heuristics, several RL models achieve better solutions, i.e., they provide higher sum rates at lower costs. This indicates that the RL agents learn to selectively acquire higher-value RISs, whereas the heuristics tend to bid more aggressively and thus increase the overall cost. Third, the bid intensity parameter β effectively regulates the aggressiveness of bidding: larger β values correspond to more conservative behavior, resulting in lower costs but also reduced performance. Consequently, β provides a practical mechanism to balance the sum rate and the cost.

Fig. 6 shows the impact of the bid intensity parameter β on the bidding behavior of the RL agents. The left vertical axis indicates the normalized bid values, as defined in (13), with a maximum value of one. As β increases, the average bid value also increases, meaning that the agents become more selective and tend to focus on high-value RISs. For small values of

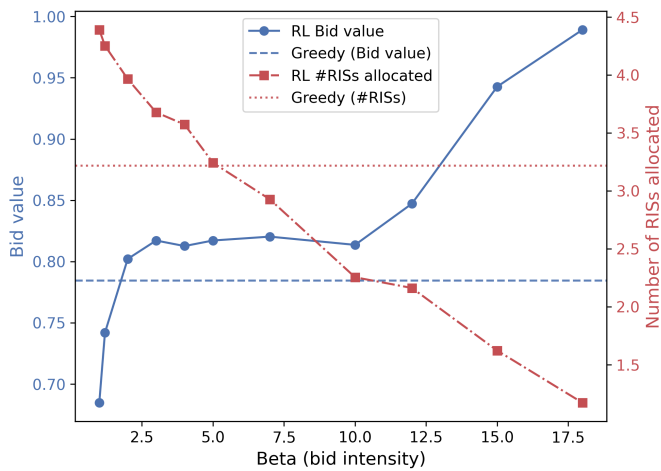


Fig. 6. Impact of the bid intensity (β) parameter on RL bidding behavior. Larger β values make agents more selective, resulting in higher average bid values but fewer allocated RISs, while smaller β values lead to broader bidding on lower-value RISs.

β , the agents are willing to also acquire lower-value RISs, whereas for medium and large values of β , the bidding is concentrated on RISs of higher utility. The right vertical axis depicts the number of RISs allocated. Here, the opposite trend can be observed: larger β values lead to fewer allocated RISs, since the agents adopt a more conservative strategy and place fewer bids overall.

VI. CONCLUSION

In this work, we investigated auction-based allocation of reconfigurable intelligent surfaces in a two-cell scenario with clustered deployments at the cell edge. Our results demonstrate that RISs substantially enhance network performance compared to scenarios without RIS assistance. Furthermore, reinforcement learning based bidding strategies outperform heuristic approaches by achieving superior trade-offs, i.e., higher sum rates at lower costs. Finally, we showed that the bid intensity parameter β provides an effective control mechanism for adjusting the aggressiveness of bidding, thereby enabling a flexible balance between spectral efficiency and expenditure. These findings highlight the potential of reinforcement learning for efficient and adaptive resource allocation in RIS-assisted networks.

REFERENCES

- [1] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, "Reconfigurable intelligent surface (RIS) assisted wireless coverage extension: RIS orientation and location optimization," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 269–273, Jan. 2021.
- [2] M. A. Msleh, F. Heliot, and R. Tafazolli, "Ergodic capacity analysis of reconfigurable intelligent surface assisted MIMO systems over Rayleigh-Rician channels," *IEEE Commun. Lett.*, vol. 27, no. 1, pp. 75–79, Jan. 2023.
- [3] Q. N. Le, V.-D. Nguyen, O. A. Dobre, and R. Zhao, "Energy efficiency maximization in RIS-aided cell-free network with limited backhaul," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 1974–1978, Jun. 2021.
- [4] N. Nisan and A. Ronen, "Algorithmic mechanism design," *Games Econ. Behav.*, vol. 35, no. 1, pp. 166–196, 2001.

- [5] P. Cramton and A. Ockenfels, "The German 4G Spectrum Auction," *The Economic Journal*, 127 (October), F305–F324, 2017.
- [6] P. Milgrom, "Putting Auction Theory to Work", Stanford University, California, Cambridge University Press, 2004.
- [7] S. Schwarz, "Gambling on Reconfigurable Intelligent Surfaces," *IEEE Communications Letters*, vol. 28, no. 4, pp. 957–961, April 2024.
- [8] C. A. Balanis, "Antenna Theory", Hoboken, New Jersey, U.S.: John Wiley & Sons, Inc., 2005.
- [9] T. Roughgarden, *Twenty Lectures Algorithmic Game Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
- [10] M. Towers et al. (Mar. 2025). *Gymnasium*. [Online]. Available: <https://zenodo.org/records/14983111>
- [11] J. K. Terry et al., "PettingZoo: Gym for multi-agent reinforcement learning," 2021, arXiv:2009.14471.
- [12] J. K. Terry, B. Black, and A. Hari, "SuperSuit: Simple microwrappers for reinforcement learning environments," 2020, arXiv:2008.08932.
- [13] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: <http://jmlr.org/papers/v22/20-1364.html>
- [14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, arXiv:1707.06347.
- [15] M. Toumi, A. Aijaz, "System Performance Insights into Design of RIS-assisted Smart Radio Environments for 6G," 2021, arXiv:2101.01102.