Reinforcement Learning-Driven Local Transactive Energy Market for Distributed Energy Resources

Steven Zhang[a], Daniel May[a], Mustafa Gül[b], Petr Musilek[a,c,*]

[a] *Electrical and Computer Engineering, University of Alberta, Canada*
[b] *Civil and Environmental Engineering, University of Alberta, Canada*
[c] *Applied Cybernetics, University of Hradec Králové, Czech Republic*

Abstract

Local energy markets are emerging as a tool for coordinating generation, supply, storage, and consumption of energy from distributed resources. In combination with automation, they promise to provide an effective energy management framework that is fair and brings system-level savings. The cooperative-competitive nature of energy trading calls for automation based on multi-agent systems with learning energy-trading agents. However, depending on the dynamics of the agent-environment interaction, this approach may yield unintended behavior of market participants. Thus, the design of local energy market suitable for reinforcement learning agents must take into account this interplay. This article introduces autonomous local energy exchange as an experimental framework based on multi-agent approach combined with a double auction market mechanism. In this setting, participants determine their internal price signals and make energy management decisions through market interactions, rather than relying on predetermined external price signals. The market properties suitable for use with reinforcement learning are determined through simulation experiments. The results show that market truthfulness is required to maintain demand-response functionality, while weak budget balancing is needed to provide a strong reinforcement signal for the learning agents. The ability of the market to facilitate demand response is illustrated by comparing the resulting behavior with two baselines: net billing and time-of-use rates. The market-based pricing was significantly more responsive to fluctuations in the community net load. In addition, for the test community, the more accurate accounting of renewable energy usage reduced bills by a median 38.8% compared to net billing.

Keywords: Transactive energy, Demand response, Distributed energy resources (DER), DER integration, Local energy market, Reinforcement learning

1. Introduction

Demand response (DR) techniques have become popular means to increase the value of distributed energy resources (DER), such as rooftop solar, while mitigating

*Corresponding author

the negative effects of their intermittent nature. DR methods can be direct or indirect. Indirect DR aims to change customer behavior using an incentive signal, usually through monetary means [1, 2, 3]. Direct DR grants grid operators immediate control to perform grid balancing. As DER adoption continues, centralized approaches to DR will encounter scalability barriers [4, 1, 5, 6], and more granular and robust control will be necessary due to the increased intermittency and stochasticity of supply and demand. Despite the effort to tackle some of these challenges [7, 8, 9], it is increasingly clear that alternative, decentralized solutions, must be explored. In this context, transactive energy (TE) is gaining popularity as a design framework for decentralized DR [5, 6]. The U.S. Department of Energy Gridwise Architecture Council defined TE as "a system of economic and control mechanisms that allows the dynamic balance of supply and demand across the entire electrical infrastructure using value as a key operational parameter" [10]. A locally constrained TE system is often referred to as local energy market (LEM). In lieu of a formal definition of LEM [11, 12], we adopt the description developed by Mengelkamp et al. [11], i.e. "a market platform for trading locally generated (renewable) energy among residential customers within a geographically and socially close community. Supply security is ensured through connections to a superimposed energy system."

High DR participation rates, especially within a LEM, can only be maintained with automation. Expert-designed, rule based systems have initially been considered for this purpose. Learning-based approaches such are now preferred, mainly because of their robustness and scalability. However, the vast majority of existing approaches that apply learning methods to LEM do not tailor the market mechanism to the algorithm used for automation. This is especially problematic given the fact that most established LEM mechanisms were designed for human participants or rule-based system automation [13]. This results in suboptimal DER utilization, as the LEM is not appropriately adjusted to best leverage the potential of the automation approach used. This is exacerbated for reinforcement learning (RL) agents that can quickly learn to exploit loopholes in competitive-collaborative multi-agent settings [14]. Mengelkamp et al. clearly identify this research gap in their review, stating that "a comprehensive comparison of the impacts of different trading designs (especially market mechanisms) should be carried out. Specifically, the impact of different allocation mechanisms on the market objectives and agent behavior need to be evaluated" [11]. Their later work [15] follows the same reasoning, noting that both agent design and market design influence the resulting system behavior. We argue, that the LEM mechanism should be tailored to adequately empower the strengths of their automation methods and to mitigate their potential weaknesses.

To the best of our knowledge, there has been no contribution that explicitly designs experiments to identify the requirements that a specific LEM market mechanism must fulfill to be well compatible with RL actors. This article aims to provide a starting point to fill this research gap. We narrowly focus on the following two questions:

- What are the required properties of LEM settlement mechanism suitable for deployment of RL-based automation?
- Does the resulting market behavior effectively support DR for LEM with high

penetration of DER?

To answer these questions, three different settlement mechanisms are examined that cover a set of established criteria for auction environments. The most suitable market design is found by analyzing the agent policies developed for each mechanism. This study then models the resulting LEM transactions as a dynamic price signal and compares its economic performance with existing pricing methods.

This article is organized in five sections. Section 2 provides the necessary background and describes the related work. Section 3 introduces the proposed autonomous local energy exchange (ALEX) and describes it as a stochastic game. Section 4 describes two sets of experiments. The first set is designed to identify settlement mechanism suitable for market automation using learning agents. The second set performs an economic analysis of the selected mechanism and compares its performance with several benchmarks. Major conclusions are summarized in Section 5, along with possible directions for future work. Three appendices describe the principles of net billing (Appendix A), transactive energy simulator T-REX (Appendix B), and details of the specific market design used in this article (Appendix C).

2. Background and Related Work

This section provides a brief introduction to RL and more in-depth review of the related work, focusing on articles that combine RL with LEM. For a broader context of LEMs, the reader may refer to a general review by Mengelkamp et al. [11], game-theory focused review by Pilz et al. [12], and review of LEM settlement and market mechanisms by Khorasany et al. [13].

2.1. Reinforcement Learning

There have been numerous approaches used to address the optimization problems inherent to demand side management [1, 4, 5, 6], including mixed-integer programming, stochastic programming, and dynamic programming. After RL demonstrated great competence in partially observable, stochastic game environments [16, 17, 18], this model-free, sequence-oriented, semi-supervised machine learning framework has also gained popularity as a control method for DR [19]. The learned policy can substitute the solution of the equivalent optimization problem at each time step. As a result, RL approaches are more computationally efficient at scale, when compared to conventional optimization methods.

In the RL setting, illustrated in Fig. 1, an agent learns to maximize the return G by interacting with its environment through actions a while receiving observations of the environmental state s. G is commonly defined as the expected, discounted cumulant of future reward R

$$G_t = \sum_{i=t}^{T} \gamma^{(i-t)} R_i, \ \forall \gamma \in [0...1], \tag{1}$$

where γ is the discount factor. The expected value of G_t, given state s_t or the state-action tuple (s_t, a_t), is referred to as the state value

$$V(s_t) = \mathbb{E}\big(G_t|_{s_t,\pi(s_t)}\big), \tag{2}$$

or action value

$$Q(s_t, a_t) = \mathbb{E}\big(G_t|_{s_t,a_t}\big), \tag{3}$$

respectively. An RL agent acts according to a policy

$$\pi : A \times S \rightarrow [0...1]., \tag{4}$$

Policy is a (probabilistic) mapping of the state space S on action space A

$$\sum_a \pi(a, s_t) = 1. \tag{5}$$

This allows the definition of the state value V as action value Q weighted by π

$$V(s_t) = \frac{1}{n_a} \sum_a \pi(a, s_t) Q(s_t, a). \tag{6}$$

This system of equations (1-6) is sufficient to broadly classify all RL algorithms along two axes: the learned function and the relation between the target and behavior policy. According to the learned function, RL algorithms can be classified as policy-gradient methods and value-based methods. The policy-gradient methods directly learn the policy π, while the value-based methods learn estimations for either V or Q, and employ a fixed mapping of these values to π. RL algorithms can also be classified into on-policy and off-policy methods, by comparing their target and exploratory behavior. An on-policy RL algorithm explores the environment with the same policy that is optimized, while an off-policy algorithm explores the environment with a behavioral policy $b \neq \pi$.

Internally, RL algorithms often employ function approximation techniques to per-

Figure 1: Reinforcement Learning Setting

form the mapping of the state space S to the learned target, and therefore the return G. As a framework, RL is independent of the choice of state estimator. Currently, a very popular choice is the use of deep artificial neural networks. Historically, other function approximation techniques have also been used, such as tabular encoding.

The algorithm used in this paper is Q-learning, a well-established, value-based, off-policy algorithm. It learns the greedy policy, a deterministic policy that always picks a corresponding to the largest Q, by following an arbitrary annealed behavioral policy b. A popular choice for b is the ϵ-greedy policy, which takes a random action with probability ϵ and otherwise follows the greedy policy. The corresponding learning rule can be written as

$$Q^{\text{updated}}(a_t, s_t) \leftarrow (1-\alpha)Q(a_t, s_t) + \alpha\Big(R_t + \gamma \max_a\big(Q(a, s_{t+1})\big)\Big), \tag{7}$$

where α is the learning rate.

Q-learning is a relatively well-understood RL algorithm, with strong convergence criteria for the tabular function approximation case, as long as both ϵ and α are annealed towards 0 at infinity. It is also the most common algorithm in the related literature reviewed in section 2.2.

2.2. Reinforcement Learning for Local Energy Markets

Several authors investigate the combination of RL and dynamic pricing for centralized control. Notably, Kim et al. [2], and Lu et al. [3] develop RL-based approaches for dynamic pricing from the perspective of a service retailer. Both articles address difficulties of predicting participant response to a pricing schedule by mitigating the reliance on accurate customer side information. A Markov decision process is formulated based on customer behaviour models and preferences. A Q-learning agent is trained to simultaneously minimize customer costs and maximize the service provider benefit. The two approaches differ in the formulation of the reward function, which is a major influencing factor in RL algorithms, in general. Lu et al. [3] use a weighted sum of retailer and customers, while Kim et al. [2] use a modelled utility function. Although both proposed approaches successfully implement dynamic pricing strategies without scheduling, they still rely on modeling consumer behavior and preferences via utility functions.

Zhang et al. [20] train an RL agent to manage a community-shared battery and trade its resources on a TE market to maximize economy. Thus, the reward function is the economic performance of the battery. The authors show that positive economic benefits can be achieved, even when considering the running costs of the battery.

Xiao et al. [21] investigate optimized trading between a large number of interconnected microgrids using a deep Q-network (DQN) based RL agent. Similarly to Kim et al. [2], a utility function is used to determine the reward. As expected, the more sophisticated DQN algorithm outperforms the benchmark hotbooting Q-learning algorithm.

Foruzan et al. [22] investigate the behavior of self-interested Q-learning agents, exchanging energy within a microgrid through a LEM. The agent's goal is to maximize its own profit. Managed DERs include battery energy storage systems, rooftop solar, wind and diesel generators. The participants' stochastic behavior is approximated using random models. The authors investigate several micro grid configurations and perform an in-depth hyperparameter study of the RL algorithm with respect to return, self-sufficiency and fairness.

Zhou et al. [23] combine a fuzzy rule-based system with Q-learning to train agents to exchange energy resources over a peer-to-peer LEM setup whose pricing is directly tied to the ratio of supply and demand. The authors investigate the performance of several community configurations with ranging number of battery energy storage systems and renewable generation assets. They show that such a system setup generally achieves lower bills than TOU and net-billing baselines.

Chen et al. [24] employ a DQN variant to automate the interactions of prosumers equipped with battery energy storage system in a LEM. The RL agents' action space consists of four distinct, discrete actions covering buy/sell and charge/discharge operations. The learned policy surpasses an intuitive, rule-based strategy. It also outperforms a pure random policy equivalent to a zero-intelligence agent, originally proposed by Ghode et al. [25] as a baseline for agent competence in automated markets. In another article, Chen et al. [26] investigate the function of Q-learning based energy brokers as LEM consensus mechanism for settlements with profit used as the agent's reward. Using several ablation and sensitivity studies, the authors show that the brokers efficiently learn how to maximize their own profit and the efficiency of the market.

Kim et al. [27] extend a DQN variant designed for stock trading applications to manage participation of a household in peer-to-peer energy exchange. Using experiments with different rate schemes, they demonstrate that the developed agent outperforms the simplified versions in terms of loss minimization and revenue maximization.

Bose et al. [28] focus on emerging participant interaction within a fixed LEM setup under differing levels of DER penetration. The authors demonstrate that RL-based agents in such environment can cause partial energy self-sufficiency to emerge. They also show that the degree of self-sufficiency and the complexity of agent interactions depends on the level of DER penetration within the market.

Mengelkamp et al. [29] study three different extensions of the Erev-Roth RL algorithm applied to automate LEM participation. They find that the extensions further increase the self-sufficiency of the LEM when compared to the original Erev-Roth algorithm [30].

Mengelkamp et al. [15] compare a peer-to-peer LEM against a closed book, double-auction LEM with settlement rounds. The authors compare the performance of zero-intelligence agents and "intelligent" agents adopted from Nicolaisen et al. [31] on both LEM designs. They show that all market scenarios offer similar economic advantages, with the peer-to-peer LEM used by intelligent agents slightly outperforming the other variants. However, they also note that using one strategy on different markets results in different price trends, The authors eventually conclude that agent strategy and

market design need to be co-developed to guarantee the system's performance.

3. ALEX: Autonomous Local Energy Exchange

As shown in Section 2, the need to model customer behaviour via utility functions and heavy reliance on forecasts may hinder the robustness and scalability of traditional DR techniques. Furthermore, due to the amount of DERs that are expected to be on-line in the near future, certain infrastructure requirements may pose additional barriers for the effective deployment of DR systems. For example, the communication and computational costs for centralized control may grow too expensive, especially for high temporal resolutions necessary to fully capture the intermittent behaviour of rooftop solar panels and stochastic use patterns of electric vehicles.

To avoid these difficulties, this article proposes a distributed, multi-agent approach combined with a double auction market mechanism. When designing this approach, dubbed ALEX (autonomous local energy exchange), the following assumptions have been made:

- Participants are self-interested and, therefore, prioritize their own economic well-being in the decision making process.
- Participants are willing to defer some decision making regarding interactions with indirect DR measures to automation (e.g., using RL agents).
- Each participating unit is equipped with a smart meter, and a sufficient amount of high-resolution historical data is available to train the RL agents.
- The large-scale electricity grid that customers are connected to is an infinite bus.

3.1. Core Concept

Conceptually, ALEX is a behind-the-meter DR technique for a localized community using a double auction market as a coordination mechanism. Market participants are the customers who live within the community. However, this could be expanded to include entities that are only temporarily present, such as electric vehicles. The market employs double auctions with a fixed settlement frequency Δt. For each interval $[t, t + \Delta t)$, participants can communicate their intention to trade energy by submitting bids

$$\text{bid}_t = (q_t^{\text{bid}}, p_t^{\text{bid}}), q \in [0...q_{\text{max}}^{\text{ask}}], p \in [p_{\text{min}}, ...p_{\text{max}}], \tag{8}$$

and asks

$$\text{ask}_t = (q_t^{\text{ask}}, p_t^{\text{ask}}), q \in [0...q_{\text{max}}^{\text{ask}}], p \in [p_{\text{min}}, ...p_{\text{max}}], \tag{9}$$

where bid_t and ask_t communicate the intention to buy or sell energy, respectively. They are represented by tuples consisting of the desired quantity q and desired price p of energy to be exchanged. Quantities are expressed in Wh, and can range from 0 to a designated maximum. $q_{\text{max}}^{\text{ask}}$ should be set to accommodate expected maximum

generation derived from historical data. Likewise, $q_{\max}^{\text{bid}}$ should be set to accommodate the expected maximum load demand. Similarly, prices in $ are within a designated window between $p_{\min}$ and $p_{\max}$. Bids and asks are settled pairwise at the end of each settlement round, returning a settlement signal m_t to each participant

$$m_t = (q_t^{\text{settlement}}, p_t^{\text{settlement}}) \tag{10}$$

More details on the market implementation are provided in Appendix C. It is important to stress that in this setting, participants both determine the price signals and make energy management decisions through market interactions, whereas most other DR approaches simply have agents react to external price signals.

The settlement signal m_t is represented as a list of tuples containing the settled quantities, and the respective prices. It is important to note that participants only receive information about their settlements and, therefore, do not have access to information on the behaviour of other participants.

After the internal trades are concluded, any excess generation/demand within the community is exchanged with the electricity grid at retail prices. In this article, we assume net billing, a commonly used practice where excess energy is sold to the grid at price $p_{\text{sell}}^{\text{grid}}$ and deficient energy is purchased from the grid for price $p_{\text{buy}}^{\text{grid}}$ that includes fees. The behind-the-meter setup grants the community a window of profitability by deferring fees. This naturally bounds the range of internal market prices as follows

$$p_{\min} = p_{\text{sell}}^{\text{grid}} <= p_{\text{market}} <= p_{\text{buy}}^{\text{grid}} = p_{\max}, \tag{11}$$

If the interactions between participants are dominated by the law of supply and demand, then ALEX functions as a decentralized, indirect DR tool. Its pricing schedule is strongly correlated with the ratio of supply and demand within the market. This naturally provides economic incentives for all market participants to balance supply and demand. We demonstrate this using the experiments described in Section 4.1.

3.2. ALEX as a Stochastic Game

To analyze the properties of the proposed approach, the auction and strategic bidding by the actors can be described using a suitable mathematical model. A game-theoretic representation of ALEX can be derived by modelling the interactions of participants as a discounted stochastic game

$$\Gamma \coloneqq (n, L, S, A, P, R) \;\; \forall t \in [0...T], \lambda \in (0...1), \tag{12}$$

where n is the number of players, L is the list players of length $|L| = n$, S is the state space, A is the action space, P represents the state transition probabilities, R is the reward function, t is the current time step over the modelling period $[0...T]$, and λ is the discount factor.

Both S and A can be decomposed into n individual components S^i and A^i, as shown below

$$S = S^1 \times ... \times S^n, \tag{13}$$

$$A = A^1 \times ... \times A^n. \tag{14}$$

Superscript i refers to a specific individual L^i, while the subscript is reserved for time t. Note that the action space A is separated in notation from a specific set of actions a_t at time step t, as in the commonly used RL nomenclature introduced in Section 2.

State transition probabilities are defined for any set of actions a_t taken at time step t, as follows

$$\forall a_t : \ P(S_{t+1}|S_t, a_t) \coloneqq S_t \to S_{t+1} \tag{15}$$

Analogous to the RL setting, the reward or payoff in the stochastic game at time step t is defined by

$$R_t \coloneqq S \times A \to r, \tag{16}$$

which maps from (S_t, a_t) to a real number $r \in \mathbb{R}$. Similarly, each agent aims to maximize their own return G_t (1). Thus, all participants use their individually developed policy π^i (4 and 5), to determine action set a_t^i based on observations from S_t^i.

At each time step, all agents can interact with the market by submitting bids (8) and asks (9). This leads to the following definition of action

$$a_t^i = (\text{bid}_t^i, \text{ask}_t^i, e_t^i), \tag{17}$$

where the additional parameter, e_t^i, is reserved for future expansion of the model, e.g., to define nonmarket actions, such as battery management or thermal load control.

Finally, the state observations for each agent are defined as follows

$$S^i = (d_t^i, g_t^i, m_{t-1}^i), \tag{18}$$

where d_t^i and g_t^i are, respectively, the load demand and generation at time t, and m_{t-1}^i are settlements received at time $t-1$.

Note that the transition probabilities P result from the collective actions of all agents. However, due to the pairwise settlement mechanism, market design, and the observation space, P is not fully accessible to L^i. This ensures that the developed model is a truly stochastic game.

At least one stable Nash equilibrium is guaranteed to exist within Γ, as long as n, S and A are finite. This condition can be guaranteed by limiting prices p to a reasonable decimal place accuracy (e.g., 4 or 5 significant digits commonly used in banking). A is logically bounded by the condition previously defined by (11). As a result, S must also be finite, and therefore each implementation of ALEX is guaranteed to exhibit at least one stable Nash equilibrium.

3.3. Automation using Reinforcement Learning

Since the interaction through the developed stochastic game Γ requires strategic competence, automating the interactions of participants (typically prosumers, but theoretically any grid-connected entity) with the LEM is a reasonable response to the difficulties of accurately modeling customer behaviour. The proposed approach

centers around training RL agents to perform market interaction and energy management actions, compensating for nonoptimal human behaviour. RL is theoretically very suitable for this task, as the stochastic game described in the previous subsection is equivalent to a Markov decision process under an established set of criteria, outlined in [32]. The framework developed in this section is set up to be algorithm agnostic. However, the subsequent experiments described in Section 4.1.1 employ independent Q-learning.

In this article, the reward function, R^i, is formulated as follows

$$\begin{aligned} R_t^i =& \text{cost}_t^{i,\text{LEM}} - \text{profit}_t^{i,\text{LEM}} + \\ &+ \text{cost}_t^{i,\text{grid}} - \text{profit}_t^{i,\text{grid}} + c, \end{aligned} \tag{19}$$

where,

$$\text{cost}_t^{i,\text{LEM}} = q_t^{i,\text{settled}-\text{bids}} \times p_t^{i,\text{settled}-\text{bids}}, \tag{20}$$

$$\text{profit}_t^{i,\text{LEM}} = q_t^{i,\text{settled}-\text{asks}} \times p_t^{i,\text{settled}-\text{asks}}, \tag{21}$$

$$\text{cost}_t^{i,\text{grid}} = q_t^{i,\text{grid}-\text{buy}} \times p_t^{i,\text{grid}-\text{buy}}, \tag{22}$$

$$\text{profit}_t^{i,\text{grid}} = q_t^{i,\text{grid}-\text{sell}} \times p_t^{i,\text{settled}-\text{sell}}, \tag{23}$$

and,

$$q_t^{i,\text{settled}-\text{bids}} + q_t^{i,\text{grid}-\text{buy}} = d_t^i \text{ (load demand)}, \tag{24}$$

$$q_t^{i,\text{settled}-\text{asks}} + q_t^{i,\text{grid}-\text{sell}} = g_t^i \text{ (generation)}. \tag{25}$$

Other considerations, such as social welfare costs, are explicitly excluded as the current aim is to study participant and system behaviour using pure economic performance. Nevertheless, these factors may be included in future studies. Since RL agents are trained using high-frequency smart meter data, explicit customer behaviour models can be omitted, as a sufficient amount of data can better capture nuanced and individualized customer behavioural patterns.

4. Experiments and Discussion

4.1. Suitability of Settlement Mechanism

Fig. 2 illustrates a typical participant, represented by a prosumer's home. The home may contain any combination of generation, storage, and controllable loads. For the experiments, the generation and load sources are taken directly from smart meter data. As the focus of this article is to study the LEM's behavior under differing ratios of available supply and demand, profile shaping via load shifting or battery storage are not considered.

4.1.1. Experimental Design

This experiment focuses on the first research question: What are the required properties of LEM settlement mechanism suitable for the deployment of RL-based

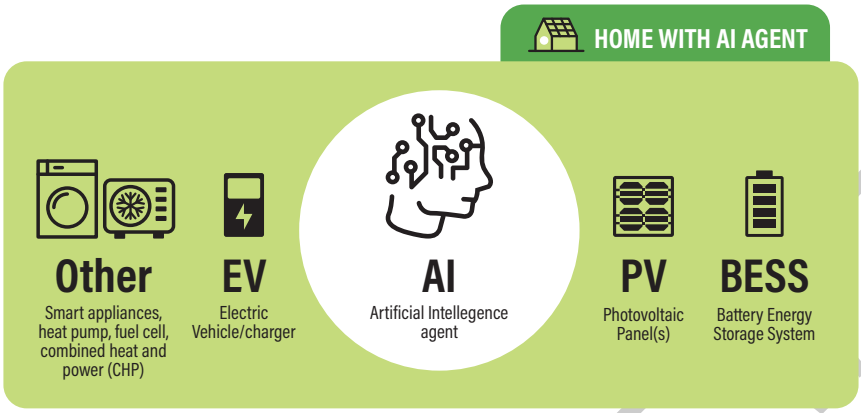

Figure 2: Diagram showing a typical participant, represented as a prosumer's home

automation? The results of this experiment will also show how different market properties influence the policies learned by the RL agents. Since the environment (i.e., the market) plays just as important role as the learning algorithm, it is imperative to establish the most suitable market mechanism for subsequent research and implementation of more complex agent designs, and action and strategy spaces. The scope of the experiment is carefully managed to magnify the influence of the settlement mechanism on the resulting agent policies while providing strong convergence bounds despite ALEX's properties as a partially observable, nonstationary environment. As a reminder, the market design used in this study and the rules of interaction are described in detail in Appendix C. In this experiment, three different settlement mechanisms with varying market properties are tested:

1. Average-Price (M1): Trades are settled if the bid price is greater than or equal to the ask price. The settlement price is the average of the bid and ask prices.
2. Exact-Match (M2): Sellers and buyers choose bid and ask prices from a list of available prices. Trades are settled if the bid price equals the ask price.
3. Exact-Price (M3): Trades are settled if the bid price is greater than or equal to the ask price. The buyer buys from the auctioneer at the bid price, and the seller sells to the auctioneer at the ask price.

Any double auction mechanism can be described by the following properties [33]: individual rationality[1], budget balancing[2], truthfulness[3], and economic efficiency[4].

[1]Individual rationality states that no participant should lose money from joining the auction

[2]There are two variants of budget balancing: weak and strong. In a weak budget balancing system, a portion of the money transferred also goes to the auctioneer; this is in addition to money transfers between participants which are the only type of exchange in a system with strong budget balancing.

[3]The dominant strategy in a truthful market is for the participants to report prices at what they believe should be the true value of the item to be exchanged.

[4]In an economically efficient system, at the end of all trading, the items should be in the hands of participants who bid the highest value.

An ideal mechanism satisfies all four properties, but it cannot be realized in practice. Since the design of ALEX and the use of RL agents ensures economic efficiency and individual rationality, the three settlement mechanisms can be differentiated by truthfulness and budget balancing alone, as shown in Table 1.

	Market Property Settings			
Mechanism	**Individual rationality**	**Budget balancing**	**Truthfulness**	**Economic efficiency**
M1	Yes	Strong	False	Yes
M2	Yes	Strong	True	Yes
M3	Yes	Weak	True	Yes

Table 1: Settlement mechanism properties

Three scenarios with different community supply/demand ratios are evaluated for all considered settlement mechanisms: over-supply (10:1), over-demand (1:10), and perfect balance (i.e. equal supply and demand). Each mechanism is evaluated based on the policies developed by the agents and the resulting market behaviour based on emerging equilibrium bid, ask, and settlement prices, given the same training curriculum. The goal is to find a market mechanism that follows the law of supply and demand, and is compatible with RL agent learning behavior. Such a mechanism is expected to produce the following results:

- Excess supply case: The generators compete for demand, driving ask prices low with bid prices following.
- Excess demand case: The consumers compete for supply, driving bid prices high with ask prices following.
- Equal supply and demand case: The bid and ask prices converge around the middle of the available price range.
- For all cases: The mean bid, ask, and settlement prices should have low spread.

A set of $n = 4$ independently learning participants is considered. Two participants with $d^i > g^i$ act as buyers, and the remaining two participants with $d^i < g^i$ act as sellers. Two of each type of participant maintain competition on both sides of the market and should prevent monopolistic behaviour. Steady-state (flat, time-invariant) energy profiles are employed for each agent, with the collective load demand and supply corresponding to the previously given ratios

$$\frac{g^{\mathrm{LEM}}}{d^{\mathrm{LEM}}} = \frac{\sum_i g^i}{\sum_i d^i}. \tag{26}$$

This setup collapses the observation space of each agent to a single point and fixes q^i_{bid} or q^i_{ask} to the residual load. This allows to further improve the purity of the experiment by learning only the price policy. From the view of a single agent,

this transforms the experiment into a partially observable, nonstationary multi-armed bandit, where the number of arms corresponds to the number of discrete price actions $|p|$. For each individual participant, an independent tabular Q-learning algorithm can be used, with ϵ-greedy exploration policy and learning rate α, as described in section 2. This maintains loose convergence guarantees despite the properties of the resulting environment [34]. For this experiment, α is set to 0.1, γ to 0.99, and ϵ to 0.98. Values of ϵ and α are annealed starting from episode 100 to balance exploration and convergence speed, with a multiplier of 0.98 per episode. Under this simple setup, if the agents fail to develop policies that reflect the previously mentioned criteria, the respective mechanisms will be considered infeasible for subsequent use.

This experiment was performed using the T-REX simulator, which is described in Appendix B. The simulations were run on a workstation with Ryzen 9 3900X processor and 32GB of 3200MHz DDR4 memory. In this specific setup, each episode took approximately 5 minutes. Detailed experimental configurations can be found on the GitHub repository of the project [35].

4.1.2. Results and Discussion

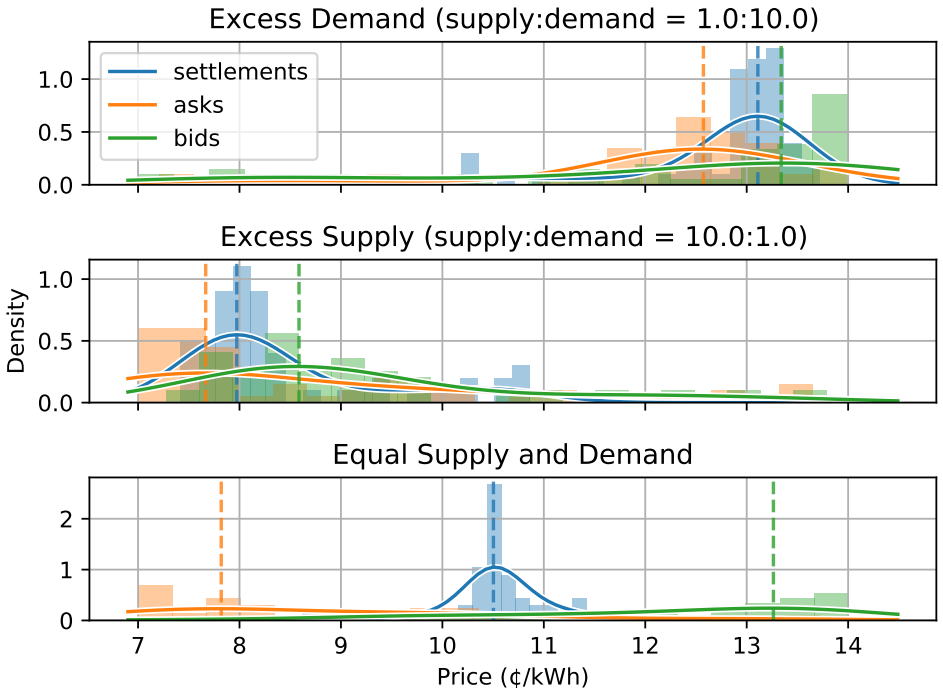

Figure 3: Validation policies for bid, ask, and resulting settlement prices for agents operating under M1 for episodes 70 to 100. Discrete policies are shown by the histograms. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters. Means of density plots are shown by dashed lines.

Recall that, in ALEX, participants both determine the price signals and make energy management decisions through market interactions. Therefore, it is important

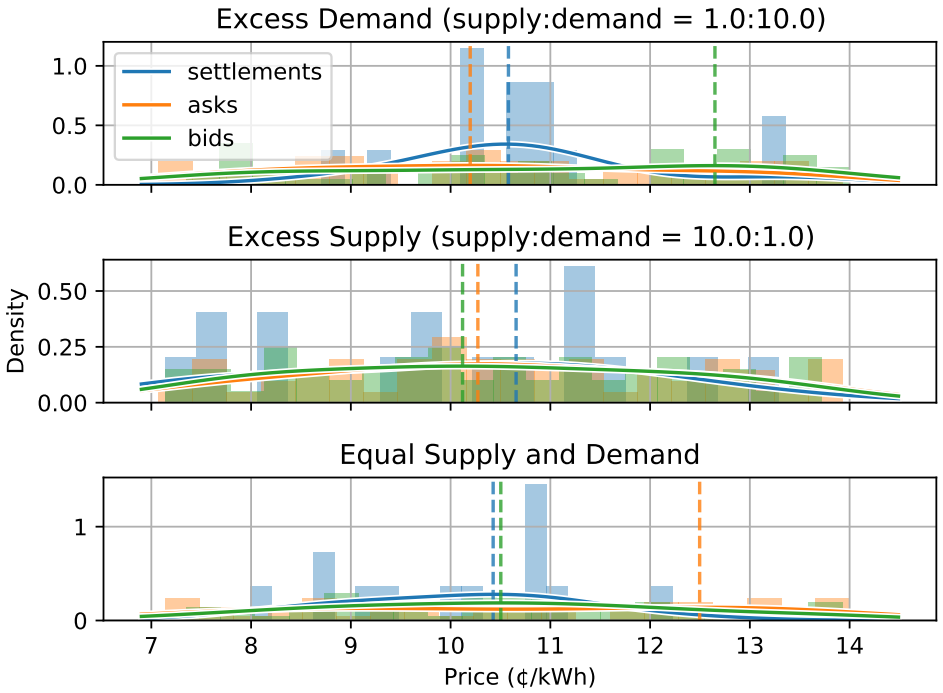

Figure 4: Validation policies for bid, ask, and resulting settlement prices for agents operating under M2 for episodes 70 to 100. Discrete policies are shown by the histograms. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters. Means of density plots are shown by dashed lines.

to study the agent actions as well as the resulting settlement prices. Fig. 3 shows, for settlement mechanism M1, the policies learned for bid and ask prices, as well as the resulting settlement prices as density plots. The bid, ask, and settlement prices are, in general, closely clustered on the expected side of the price range for both unbalanced cases. However, the balanced case reveals a critical problem. While the settlement prices are concentrated in the middle of the price range (as expected), both bid and ask prices diverge to near the extremities. This phenomenon results from the lack of truthfulness of M1. Agents have no incentive to submit the bid and ask prices that correspond to what they believe should be the value of energy (near the settlement price). Since M1 calculates the settlement price as the average of each pair of bid/ask, this strategy increases the chance of reaching a settlement. However, at the same time, it also increases the reward if an opponent follows a truthful strategy. This behavior is evidently the optimal strategy to employ in this scenario. However, the price divergence is problematic, especially when continuous, unbounded action spaces were used for price selection: exceedingly large bid/ask prices may cause settlement prices to become unstable. Because of this risk, M1 is disqualified.

Fig. 4 shows the results for settlement mechanism M2. Unlike the previous mechanism, M2 shows no clear convergence for any case. A possible explanation is that

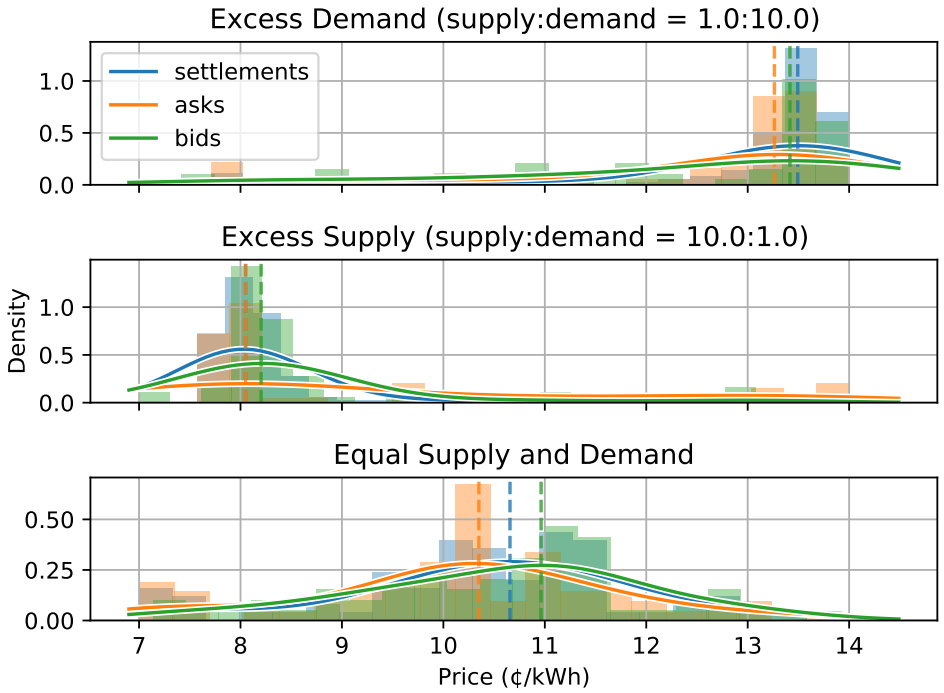

Figure 5: Validation policies for bid, ask, and resulting settlement prices for agents operating under M3 for episodes 70 to 100. Discrete policies are shown by the histograms. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters. Means of density plots are shown by dashed lines.

M2 satisfies the conditions for an ideal double auction market, which is known to be impossible to practically implement according to the Myerson–Satterthwaite theorem [33]. Another possibility is that strong budget balancing drastically decreases the number of successful settlements, which results in sparse rewards within this nonstationary environment. Therefore, despite the theoretical existence of a Nash equilibrium, agents are unlikely to discover it due to the lack of feedback. While it may be possible for M2 to converge, given a sufficiently long training time, the fact that it fails to show convergence using the same simulation parameters as M1 and M3 makes it less desirable. RL agents are often expected to update policies on data streams in real-time, which is much slower than in simulations. The real-time equivalent of the training steps is 200 weeks, or approximately four years of one-minute resolution data. However, there is still a lack of historical smart meter data. Therefore, if this system were deployed in a real environment, it would have to learn in real-time. As a result, the market mechanism M2 is disqualified as unsuitable for real world applications.

Fig. 5 shows the results for settlement mechanism M3. Similar to M1, the bid, ask, and settlement prices are closely clustered on the expected side of the price range for both unbalanced cases, even more closely together. However, unlike M1, the balanced case shows similar behaviour, with prices concentrated in the middle of

the price range. Therefore, it can be concluded that M3 has the truthfulness property. The agents have little incentive to set bid/ask prices that deviate too far from the settlement prices, which should closely approximate the true value of energy for each ratio of supply-to-demand. Consequently, M3 is qualified for further research as it satisfies the previously mentioned selection criteria.

The supply-to-demand ratios used in the initial experiments were quite extreme. Examination of the settlement prices for more balanced ratios should provide a more thorough picture of market behaviour. Therefore, the ratios 1.5:1 and 1:1.5 are added to the experiments for M3. The simulations are extended by 100 episodes with annealing as described in Section 4.1.1. As shown by the results in Fig. 6, the prices settle slightly lower than the balanced case for supply-to-demand of 1.5:1, and slightly higher for supply-to-demand of 1:1.5. This confirms the dominance of the law of supply and demand, as the settlement prices follow the supply-to-demand ratio. Further experiments with more ratios of supply and demand will be performed to develop an empirical model of the price behaviour.

In summary, the experiments show that efficient RL agent training requires weak budget balancing, resulting in a stronger, denser reward signal. Truthfulness is necessary for the emerging policies to truly reflect the law of supply and demand. In a LEM with these two properties, the individual rationality of agents maximizes the value exchanged between participants, guaranteeing economic efficiency as a result of convergence. Even though perfect budget balancing has not been achieved, this may be desirable for deployment: a small profit for the auctioneer can be used to maintain the infrastructure necessary for operating the market.

4.2. Economic Study

4.2.1. Experimental Design

This experiment focuses on the second research question: Does the resulting market behavior effectively support DR for LEM with high penetration of DER? This investigation is performed by comparing the proposed approach with conventional pricing schemes, such as net billing[5] and time-of-use. As indicated by the results from the previous experiment, the prices in market equilibrium are dominated by the law of supply and demand (see Fig. 6). By performing additional simulations with alternative proportions of supply and demand, an empirical model of the price behaviour can be obtained via interpolation. Such a model can be used as a simple approach to set local market prices, without implementing an actual auction-based market. The supply-to-demand ratio for a local market can be derived from metering data.

The economic study is conducted using a residential community microgrid with ten participants. Due to the lack of suitable smart home data from Canada, energy profiles from the openly available SunDance data set [36, 37] are used. Ten energy profiles have been randomly selected to assemble the virtual community. The IDs of the selected customers are as follows: 10011, 1001625, 1002714, 10068, 100703,

[5]The distinction between net billing and net metering is clarified in Appendix Appendix A.

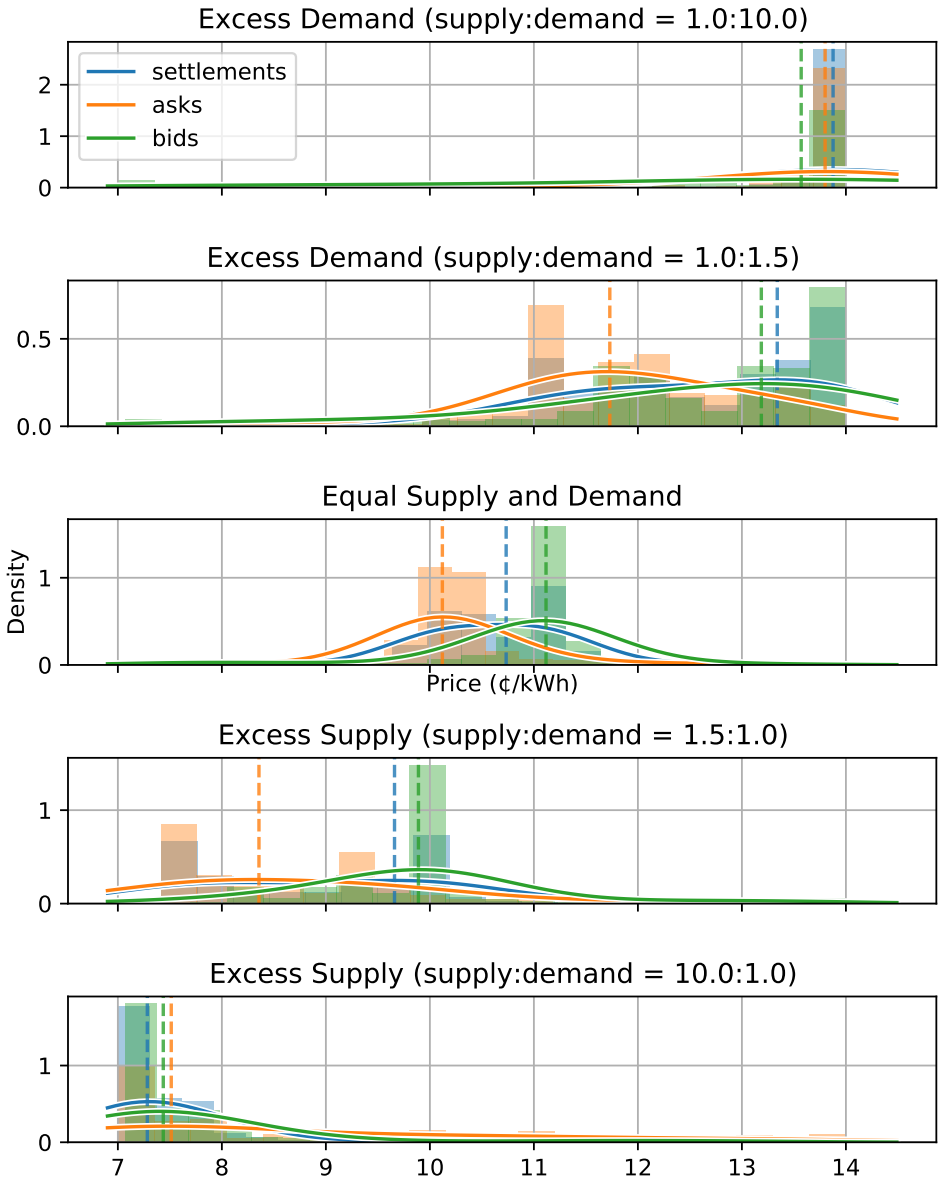

Figure 6: Validation policies for bid, ask, and resulting settlement prices for agents operating under M3 for episodes 100 to 200. Exploration factor and learning rate are annealed starting from episode 100 with a multiplier of 0.98 applied at the beginning of each episode. Discrete policies are shown by the histograms. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters. Means of density plots are shown by dashed lines.

1001420, 1003173, 1001230, 100114, 100196. All participants are prosumers participating in energy trading to gain economic benefits. The microgrid is assumed to be on a single bus behind a community smart meter. Similar to the previous experiment, no load shaping is performed. To illustrate the changes of supply-to-demand behaviour of the test community, the aggregated values of supply and demand over a single summer day (June 1, 2015) are plotted in Fig. 7.

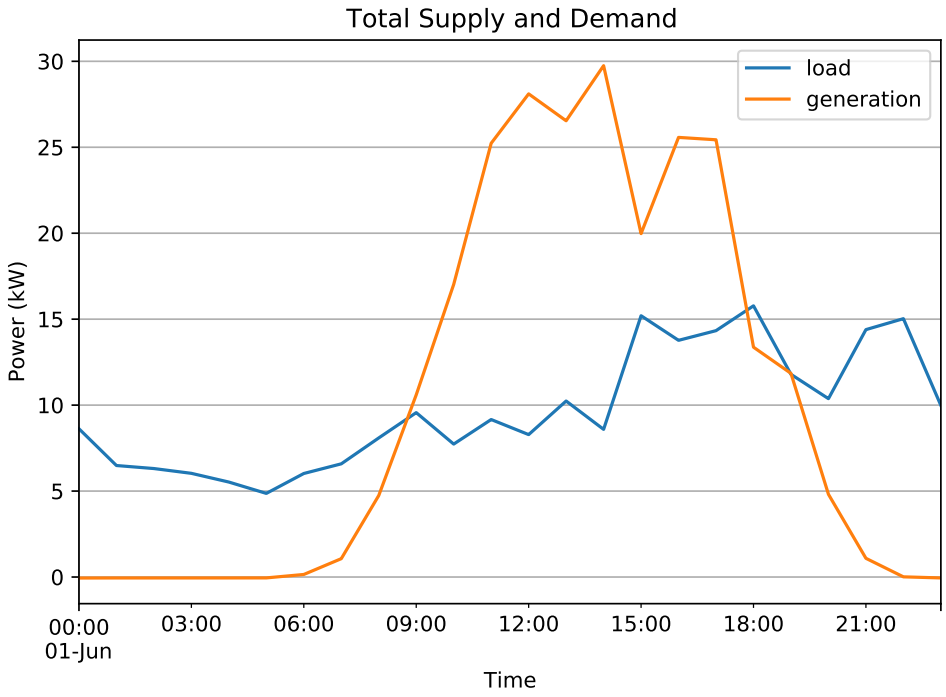

Figure 7: Total supply and demand profile of the residential community test over one summer day in June 1, 2015

The experiment evaluates the changes of electricity bills caused by enabling a local energy market that sets energy exchange prices based on the local supply and demand. Because the local market price is time-varying and supply-to-demand dependent, the LEM-based pricing can also be compared to a benchmark TOU pricing schedule.

4.2.2. Results and Discussion

The system-wide market model is developed using the data from the previous experiments supplemented by X additional demand ratios (values). The resulting pricing model is shown in Fig. 8. Note that in real settings, where the load demand curve and DER availability of each participant are unique, ALEX agents may develop personalized pricing schedules. The goal of this experiment is to evaluate the economic performance of an ALEX-based trading system and to compare it with common tariffs that do not use individual pricing schedules.

The resulting equation for the price curve is as follows:

$$P(s,d) = \begin{cases} P^{\text{grid}}_{\text{NB,load}} & P(s,d) \geq P^{\text{grid}}_{\text{NB,buy}} \\ P^{\text{grid}}_{\text{NB,gen}} & P(s,d) \leq P^{\text{grid}}_{\text{NB,sell}} \\ -0.0254\frac{s}{d} + 0.1426 C^{\text{TOU}}_{H,M,L} & \text{if buying} \\ -0.0280\frac{s}{d} + 0.1299 C^{\text{TOU}}_{H,M,L} & \text{if selling,} \end{cases} \tag{27}$$

where s is energy supply, d is energy demand, $P^{\text{grid}}_{\text{NB,load}}$ is price of electricity when

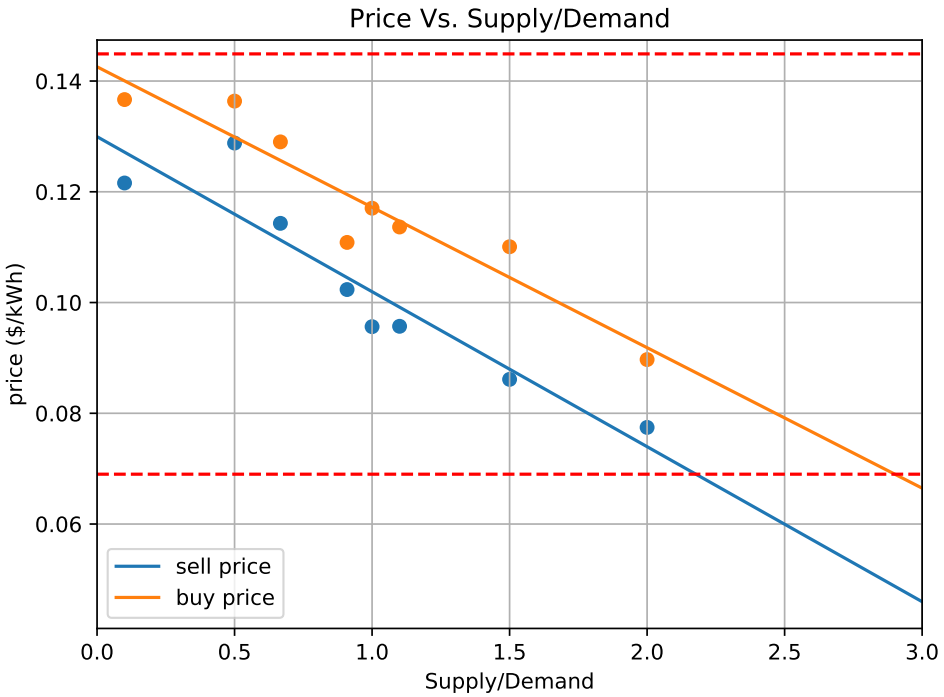

Figure 8: Pricing model developed for the test local market. The dotted lines show the price boundaries defined by (11). Linear regression between the price points leads to a well-fit, generalized mathematical model.

buying electricity from the grid under net billing, $P^{\text{grid}}_{\text{NB,gen}}$ is price of electricity when selling electricity to the grid under net billing, and $C^{\text{TOU}}_{H,M,L}$ are the adjustment factors when TOU is used.

A pricing schedule for the local market of the energy community and a selected day can be obtained by applying this model to specific energy profiles, as shown in Fig. 9 for the sample profiles from Fig. 7.

The internal price determined using the model corresponds well to the ratio of supply and demand of the community throughout the day. For example, at midnight, when solar generation is nil, the price for selling energy to a peer is \$0.1449, which is the same for the buyer as if purchasing energy from the grid. Later in the morning, at 7:00AM, when solar energy becomes available, the price for selling to peers lowers accordingly. At around 9:00AM, when generation is significantly higher than demand, the price for selling to peers drops to \$0.069, which is the same as selling to the grid under net billing.

The economic performance of all participants under this price curve is calculated and compared with net billing. The results of this comparison are shown in Fig. 10 and summarized in Table 2. As a reminder, the rules of interaction between the community participants and the grid are detailed in Appendix C. In accordance with the operating principles of net billing, described in Appendix A, the entire community is placed behind a community meter, and energy exchanged directly between peers does not incur T&D fees.

Table 2: ALEX vs. Net Billing (NB)

	Bill ($)			kWh bought from Grid			kWh bought local		kWh sold to Grid			kWh sold local	
Participant	**NB**	**ALEX**	**diff.**	**NB**	**ALEX**	**diff.**	**NB**	**ALEX**	**NB**	**ALEX**	**diff.**	**NB**	**ALEX**
10011	0.10	-0.28	383.2%	6.83	2.35	65.59%	N/A	4.48	12.88	0	100%	N/A	12.88
1001625	0.80	0.13	83.2%	10.73	2.28	78.75%	N/A	8.45	10.97	0	100%	N/A	10.97
1002714	4.34	3.26	24.8%	32.45	13.10	59.63%	N/A	19.35	5.31	0	100%	N/A	5.31
10068	2.80	1.55	44.7%	22.37	4.08	81.76%	N/A	18.29	6.41	0	100%	N/A	6.41
100703	2.35	1.32	44.0%	22.47	5.94	73.56%	N/A	16.53	13.17	0	100%	N/A	13.17
1001420	2.21	1.47	33.6%	17.63	11.96	32.16%	N/A	5.67	4.98	0	100%	N/A	4.98
1003173	7.56	5.09	32.7%	61.73	22.00	64.36%	N/A	39.73	20.03	0	100%	N/A	20.03
1001230	3.55	2.69	24.2%	24.89	8.03	67.74%	N/A	16.86	0.80	0	100%	N/A	0.80
100114	5.22	2.70	48.2%	49.55	12.65	74.47%	N/A	36.9	28.40	0	100%	N/A	28.40
1001965	6.50	4.78	26.5%	50.78	19.07	62.45%	N/A	31.71	12.45	0	100%	N/A	12.45
Total	**35.43**	**22.70**	**35.9%**	**299.44**	**101.46**	**66.12%**		**198**	**115.40**	**0**	**100%**		**115.40**

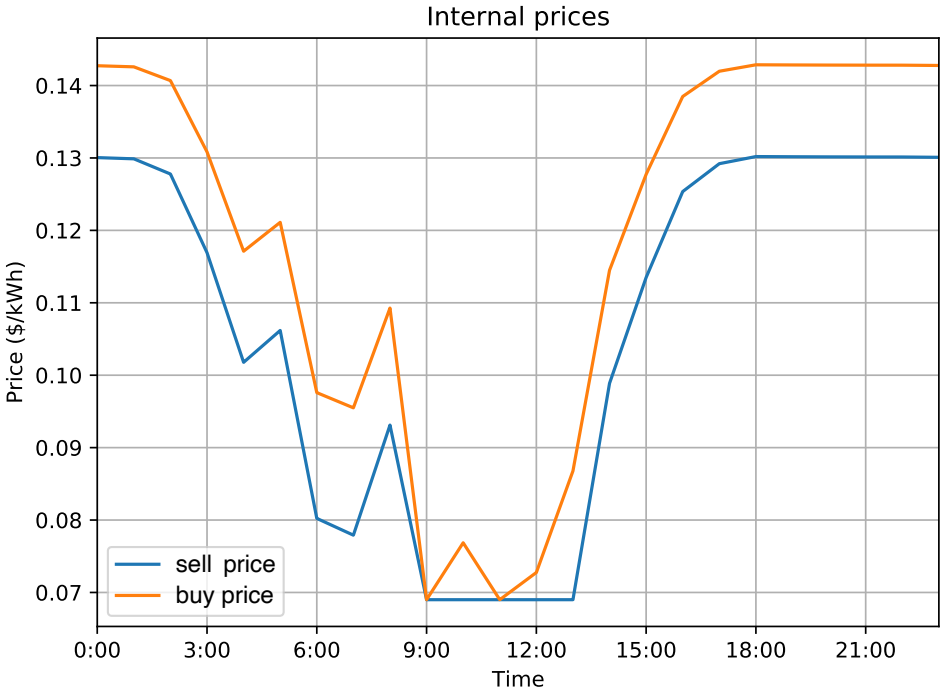

Figure 9: Internal prices of ALEX used to conduct transactions

The results in Table 2 show that the implementation of a community market can financially benefit the community as a whole, reducing the total community bill by 35.9%. The mean and median of individual bill reductions are 74.51% and 38.8%, respectively. By putting the whole community behind-the-meter, financial benefits may even be gained by those who cannot afford the expense of acquiring and installing their own DER. This is because they have direct access to excess generation of their neighbours that may be lower priced in comparison to buying from the grid. Similarly, there is an inherent financial incentive to sell excess generation to peers first, as the profits can be higher than selling directly to the grid. In other words, this setup may further socialize the benefits of DERs.

As mentioned before, the local market price is time-varying and supply-to-demand dependent. This is similar to the philosophy behind the development of TOU, which uses the time-varying supply/demand of the entire grid instead of focusing on any specific area. Therefore, these two approaches are compared to quantify their relative performance. Ontario TOU is used as a benchmark in Canada and is often referenced by utility companies in jurisdictions without TOU, such as Alberta.

Fig. 11 displays the two pricing schedules, showing the stark contrast between their shapes. Whereas the local energy price decreases toward noon due to the increase in generation, TOU increases, which suggests that there is more load than generation during this period. While it is possible that this is true due to commercial and industrial loads, which do not exist in the local market, the fact remains that the TOU is not correlated with the actual balance of load and demand in the testing locale. While this disconnect may be a cause for the lack of participation in TOU

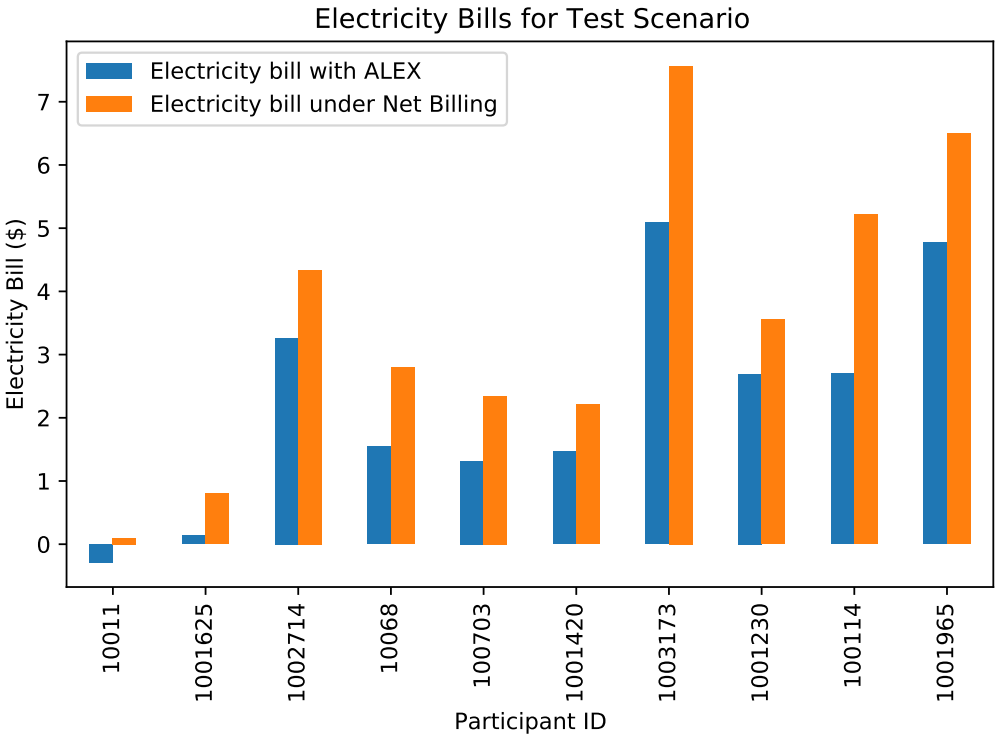

Figure 10: Electricity bill comparison between net billing and ALEX

mentioned in Section 2, it also suggests the need for very localized, highly relevant pricing signals to increase the efficiency of managing DERs and the overall system. ALEX is a highly scalable, distributed approach that generates highly relevant pricing signals at a low cost.

5. Conclusions and Future Work

DR techniques provide an effective means to manage DERs. As more such resources are installed and the energy mix becomes more complex, their management and coordination should be automated. This article explores the requirements for automating LEMs using multi-agent reinforcement learning. This exploration is facilitated by ALEX, a LEM framework that can use an arbitrary closed-book, double auction settlement system. It is used to identify the market properties that drive the policies of independent Q-learning agents to follow the law of supply and demand. After establishing an appropriate market settlement mechanism, the emergent market behaviour is compared to conventional DER integration techniques.

The first experiment trains a group of agents with three market configurations, distinguished by their general properties. The results show that truthfulness is necessary for the collective policy to reflect the law of supply and demand. The second requirement, weak budget balancing, facilitates the generation of reinforcement signals sufficient for the trading agents to learn.

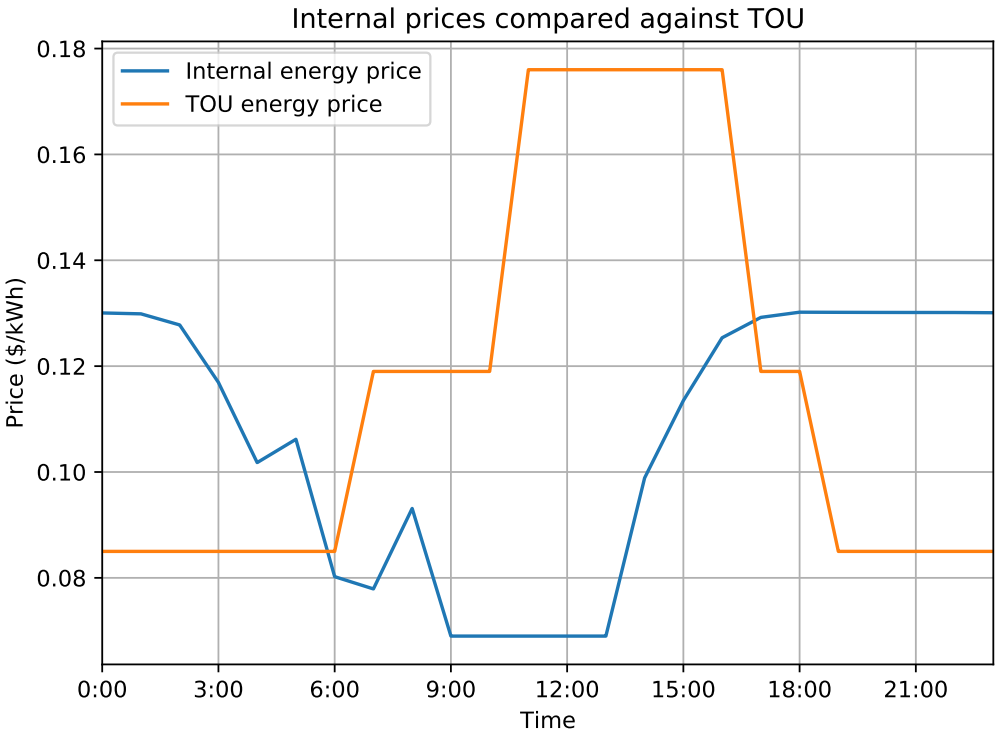

Figure 11: Local market pricing schedule compared against Ontario summer TOU prices.

The second experiment compares the resulting LEM behaviour with that of markets based on net billing and time-of-use (TOU). Since consensus pricing in ALEX strongly reflects the law of supply and demand, the resulting price signal is significantly more responsive and relevant than TOU. In terms of economic performance in the test community, the proposed approach provides a bill reduction of 38.8% compared to net billing.

The findings presented in this article lay the crucial ground for future work. We plan to investigate the integration of battery energy storage systems, along with more complex RL algorithms to increase profile shaping capabilities. This may further improve economic performance, as well as grid stability. Performance comparisons with other deployed LEM approaches will be conducted to identify the most suitable automation approach for ALEX.

References

[1] A. Fattahi, M. Deihimi, A review of demand-side management: Reconsidering theoretical framework, Renewable and Sustainable Energy Reviews 80 (2017) 367–379. doi:10.1016/j.rser.2017.05.207.

[2] B.-G. Kim, Y. Zhang, M. van der Schaar, J.-W. Lee, Dynamic pricing and energy consumption scheduling with reinforcement learning, IEEE Transactions on Smart Grid 7 (2016) 2187–2198. doi:10.1109/TSG.2015.2495145.

[3] R. Lu, S. Hong, X. Zhang, A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach, Applied Energy 220 (2018) 220–230. doi:10.1016/j.apenergy.2018.03.072.

[4] A. Meyabadi, M. Deihimi, A review of demand-side management: Reconsidering theoretical framework, Renewable and Sustainable Energy Reviews 80 (2017) 367–379. URL: https://www.sciencedirect.com/science/article/pii/S1364032117308481. doi:https://doi.org/10.1016/j.rser.2017.05.207.

[5] S. Chen, C.-C. Liu, From demand response to transactive energy: state of the art, Journal of Modern Power Systems and Clean Energy 5 (2017) 10–19.

[6] O. Abrishambaf, F. Lezama, P. Faria, Z. Vale, Towards transactive energy systems: An analysis on current trends, Energy Strategy Reviews 26 (2019) 100418.

[7] M. Yu, R. Lu, S. Hong, A real-time decision model for industrial load management in a smart grid, Applied Energy 183 (2016). doi:10.1016/j.apenergy.2016.09.021.

[8] X. Huang, S. H. Hong, Y. Li, Hour-ahead price based energy management scheme for industrial facilities, IEEE Transactions on Industrial Informatics 13 (2017) 2886–2898. doi:10.1109/TII.2017.2711648.

[9] R. de Sá Ferreira, L. A. Barroso, P. R. Lino, M. M. Carvalho, P. Valenzuela, Time-of-use tariff design under uncertainty in price-elasticities of electricity demand: A stochastic optimization approach, IEEE Transactions on Smart Grid 4 (2013) 2285–2295. doi:10.1109/TSG.2013.2241087.

[10] D. Forfia, M. Knight, R. Melton, The view from the top of the mountain: Building a community of practice with the gridwise transactive energy framework, IEEE Power and Energy Magazine 14 (2016) 25–33.

[11] E. Mengelkamp, J. Diesing, C. Weinhardt, Tracing local energy markets: A literature review:, it - Information Technology 61 (2019) 101–110. URL: https://doi.org/10.1515/itit-2019-0016. doi:doi:10.1515/itit-2019-0016.

[12] M. Pilz, L. Al-Fagih, Recent advances in local energy trading in the smart grid based on game-theoretic approaches, IEEE Transactions on Smart Grid 10 (2019) 1363–1371. doi:10.1109/TSG.2017.2764275.

[13] M. Khorasany, Y. Mishra, G. Ledwich, Market framework for local energy trading: a review of potential designs and market clearing approaches, IET Generation, Transmission & Distribution 12 (2018) 5899–5908. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-gtd.2018.5309. doi:https://doi.org/10.1049/iet-gtd.2018.5309. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-gtd.2018.5

[14] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, I. Mordatch, Emergent tool use from multi-agent autocurricula, 2020. arXiv:1909.07528.

[15] E. Mengelkamp, P. Staudt, J. Garttner, C. Weinhardt, Trading on local energy markets: A comparison of market designs and bidding strategies, in: 2017 14th International Conference on the European Energy Market (EEM), 2017, pp. 1–6.

[16] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of go with deep neural networks and tree search, Nature 529 (2016) 484–489. doi:10.1038/nature16961.

[17] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, Grandmaster level in StarCraft II using multi-agent reinforcement learning, Nature 575 (2019). Number: 7782 Publisher: Nature Publishing Group.

[18] OpenAI, :, C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, S. Zhang, Dota 2 with large scale deep reinforcement learning, 2019. arXiv:1912.06680.

[19] J. Vázquez-Canteli, Z. Nagy, Reinforcement learning for demand response: A review of algorithms and modeling techniques, Applied Energy 235 (2019) 1072–89. doi:10.1016/j.apenergy.2018.11.002.

[20] H. Zang, J. Kim, Reinforcement learning based peer-to-peer energy trade management using community energy storage in local energy market, Energies 14 (2021). URL: https://www.mdpi.com/1996-1073/14/14/4131. doi:10.3390/en14144131.

[21] L. Xiao, X. Xiao, C. Dai, M. Pengy, L. Wang, H. V. Poor, Reinforcement learning-based energy trading for microgrids, 2018. arXiv:1801.06285.

[22] E. Foruzan, L.-K. Soh, S. Asgarpoor, Reinforcement learning approach for optimal distributed energy management in a microgrid, IEEE Transactions on Power Systems 33 (2018) 5749–5758. doi:10.1109/TPWRS.2018.2823641.

[23] S. Zhou, Z. Hu, W. Gu, M. Jiang, X.-P. Zhang, Artificial intelligence based smart energy community management: A reinforcement learning approach, CSEE

Journal of Power and Energy Systems 5 (2019) 1–10. doi:10.17775/CSEEJPES.2018.00840.

[24] T. Chen, W. Su, Local energy trading behavior modeling with deep reinforcement learning, IEEE Access 6 (2018) 62806–62814. doi:10.1109/ACCESS.2018.2876652.

[25] D. K. Gode, S. Sunder, Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality, Journal of Political Economy 101 (1993) 119–137.

[26] T. Chen, W. Su, Indirect customer-to-customer energy trading with reinforcement learning, IEEE Transactions on Smart Grid 10 (2019) 4338–4348. doi:10.1109/TSG.2018.2857449.

[27] J.-G. Kim, B. Lee, Automatic p2p energy trading model based on reinforcement learning using long short-term delayed reward, Energies 13 (2020). URL: https://www.mdpi.com/1996-1073/13/20/5359.

[28] S. Bose, E. Kremers, E. M. Mengelkamp, J. Eberbach, C. Weinhardt, Reinforcement learning in local energy markets, Energy Informatics 4 (2021) 7. URL: https://doi.org/10.1186/s42162-021-00141-z. doi:10.1186/s42162-021-00141-z.

[29] E. Mengelkamp, J. Gärttner, C. Weinhardt, Intelligent agent strategies for residential customers in local electricity markets, in: Proceedings of the Ninth International Conference on Future Energy Systems, e-Energy '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 97–107.

[30] I. Erev, A. E. Roth, Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria, The American Economic Review 88 (1998) 848–881. URL: http://www.jstor.org/stable/117009.

[31] J. Nicolaisen, V. Petrov, L. Tesfatsion, Market power and efficiency in a computational electricity market with discriminatory double-auction pricing, IEEE Transactions on Evolutionary Computation 5 (2001) 504–523.

[32] F. A. Oliehoek, C. Amato, A concise introduction to decentralized POMDPs, Springer, 2016.

[33] R. B. Myerson, M. A. Satterthwaite, Efficient mechanisms for bilateral trading, Journal of Economic Theory 29 (1983) 265–281.

[34] J. Hu, M. P. Wellman, Nash q-learning for general-sum stochastic games, Journal of machine learning research 4 (2003) 1039–1069.

[35] S. Zhang, Trex-publication-resources, https://github.com/sd-zhang/publications-resources, 2021.

[36] D. Chen, D. Irwin, Sundance: Black-box behind-the-meter solar disaggregation, in: e-Energy '17: Proceedings of the Eighth International Conference on Future Energy Systems, 2017, pp. 45–55.

[37] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, J. Albrecht, Smart*: An open data set and tools for enabling research in sustainable homes, Proc. SustKDD. (2012).

[38] D. Arrachequesne, Socket.io, 2021. URL: `https://github.com/socketio/socket.io`.

[39] EPRI, Epri distribution system simulator, 2021. URL: `https://sourceforge.net/projects/electricdss/`.

[40] D. Krishnamurthy, Opendssdirect.py, `https://github.com/dss-extensions/OpenDSSDirect.py`, 2017.

[41] L. Tesfatsion, Agent-based computational economics: growing economies from the bottom up., Artificial Life 8 (2002) 55–82.

[42] D. Friedman, J. Rust, The Double Auction Market: Institutions, Theories, and Evidence, Reading, Mass., 1993.

[43] D. Friedman, A simple testable model of double auction markets, Journal of Economic Behavior & Organization (1991).

Appendix A. Net Billing

This appendix clarifies the distinction between net billing and net metering. In certain jurisdictions, such as Alberta, Canada, the electricity market is "unbundled". In simple terms, electric utilities are only in charge of building and operating the infrastructure (wires), and a multitude of retailers (which cannot be the same entity as the electric utility) are allowed to sell electricity to end users, with almost complete freedom to set the rate of electricity. Customer bills are therefore also separated into two main components: infrastructure (transmission and distribution, or T&D fees, which can have a fixed component and a variable component), and energy. Under net metering, any electricity that flows into the meter (loads) incurs both energy and variable T&D costs, and any electricity that flows out of the meter (generation) has both energy and T&D costs deducted, either as credits or cashback. Net billing is the same for loads, but only the energy component is deducted for generation. One way to avoid this infrastructure cost is to install both the solar panel and a battery behind the meter to minimize the amount of energy flowing out of the meter. The advantage of net billing is the socialized cost of infrastructure, which is more evenly divided amongst all customers. In contrast, net metering tends to shift these costs onto the segment of the population who cannot afford their own solar (this is a commonly known and often criticized problem). The disadvantage of net billing is that the return on investment (ROI) can be significantly longer due to less bill

deductions. From this perspective, net billing is a more fair baseline. It also provides more opportunities for community-based energy management, such as through local energy markets like ALEX.

Appendix B. T-REX

Appendix B.1. System Architecture

The major limitations in deploying any TE technology at a scale are communication infrastructure and computational power. This is especially true for the distribution system, where the amount of data that needs to be collected and processed for TE is orders of magnitude greater compared to the transmission system. Furthermore, the necessary infrastructure, such as SCADA, private fiber networks, voltage sensors, current sensors, etc., is typically unavailable to the distribution system and would be prohibitively expensive to retrofit.

The T-REX architecture is therefore designed around the least expensive way to implement and scale TE technology. This means that inexpensive, low bandwidth, long-range wireless mesh networks, such as LoRaWAN, can be used for reliable communications. Computing devices should also be distributed so that the total computational power of the network scales with the number of TE clients. Fig. B.12 shows the simplified architecture diagram of this approach.

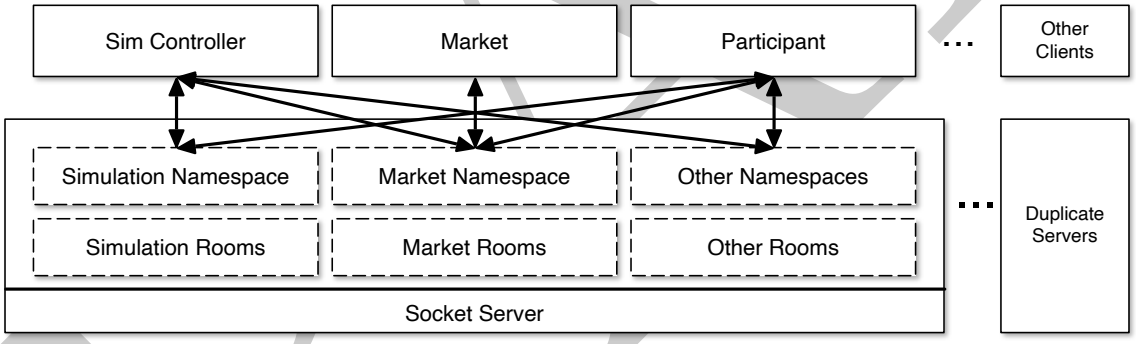

Figure B.12: Simplified T-REX V3 Architecture Diagram

Appendix B.2. Data Fabric

T-REX is built upon `socket.io` [38] as the system foundation. This guarantees compatibility, scalability, reliability, and deployability. When using T-REX in simulation mode, the asynchronous, highly parallel design provides true system-wide randomness and eliminates the need for pseudo-random sequence queues that are typically required for contemporary TE simulators. Standard networking performance and penetration testing techniques can be readily used to evaluate the performance and cybersecurity aspects of the TE systems designed with T-REX.

Appendix B.3. Clients

The functional modules of T-REX are built as `socket.io` clients. As in the deployment case, interaction between modules is facilitated by passing messages using the `socket.io` API. Although designers are free to use payloads of any permissible size, format, and endpoints, care should be taken to preserve genericity and minimize bandwidth usage. There have been three main classes of clients implemented, as shown in the architecture diagram and described below:

- Participant modules, which are in charge of energy trading and managing energy resources that are directly accessible. Participants are, for example, households and self-driving EVs.
- Non-participant modules, e.g., the TE market. The market facilitates the discovery and exchange of energy between participants.
- Simulation-only modules, e.g., the simulation controller, or a powerflow calculation module. The simulation controller augments the deployment environment to form a simulation model. It can also perform advanced functions such as training curricula for ML applications.

With a few restrictions pertaining to the simulation mode, the number of modules of each type is unlimited. The functions are also not restricted to the list described above. For example, a traffic module can function in parallel with multiple markets to guide self-driving EV participants to find optimal paths to carry passengers in conjunction with charge and discharge locations to maximize profit. Other modules that do not use traffic data to make decisions simply do not know about its existence.

Appendix B.4. Implementing TE in T-REX

TE systems can be setup in T-REX using a simple JSON configuration file. The T-REX runner assembles the modules as configured and launches them as independent processes on the assigned machine or machines at run time. Examples of configuration files can be found on the GitHub repository [35].

To launch a classical TC simulation, the following modules are required:

1. A non-participant powerflow module. The built-in implementation uses OpenDSS [39] and its Python API [40].
2. TC Market with a sub-module that generates prices based on the received powerflow data.
3. Participants containing load profiles, controllable devices, and price-reactive logic.

Fig. B.13 shows a simplified version of the sequence flow diagram of the TC co-simulation implemented in T-REX. Due to the asynchronous nature of T-REX, many independent asynchronous functions and parallel loops have been omitted from the diagram, and only an approximation of the main flow path is shown.

In the same way, T-REX can also be configured to run agent-based economics (ACE) [41] simulations with minimal modifications from the TC configuration. In the

example configurations, the only modifications are the removal of the parallel running powerflow module and swapping in the appropriate market module and agent logic submodules.

Appendix C. Double Auction Market Design for AI

Appendix C.1. Trading Mechanism

Price theory states that the price for any specific good or service is based on the balance of supply and demand. In a market-based TE approach, the role of the market is to efficiently facilitate the exchange of energy so that the price can appropriately and accurately reflect the balance of supply and demand at the time of exchange. ALEX adapts and adjusts an existing market design to fit three key considerations:

1. **Suitability for electricity grids with high penetration of DER and RES.** This means that, from a high level perspective, a market (or a collection of markets) must be able to effectively target localization and the intermittent nature of RES.
2. **Technical constraints and requirements of deployment:** Data acquisition, transportation, and cost must be minimized.
3. **Machine learning considerations for agents:** Related to the point above, ML will play an important role in trading and managing of energy resources in place of humans. For this reason, the market should be conducive to learning. One way to achieve this is to compose the market with a small set of explicit rules. The rules should provide a strong feedback signal, and they should be flexible enough to offer large action spaces.

With these considerations in mind, the final market is a modified form of double auctions [42][43]. The rules, explicitly implemented in the code, are described below:

1. It is assumed, for the time being, that the grid is an infinity bus and it can be interacted with through net billing. We therefore adapt retail electricity prices in Alberta, where buying energy from the grid costs \$0.1449/kWh, and selling earns \$0.069/kWh.
2. The local market has two energy pools: one for dispatchable sources, such as battery energy storage systems, and one for non-dispatchable sources, such as photovoltaics. This is intended to distinguish the source of energy, and to allow for the value of dispatchability to emerge.
3. Auctions settle for energy to be delivered during the one-round period from the end of the current round. However, the delivery period can be parametrically adjusted during run-time for future design explorations.
4. During the current round, participants submit bids and asks for energy to be delivered during or beyond the next delivery period.
5. A modified double auction system is used to settle trades: bids/asks are settled pairwise, with bids sorted from the highest to lowest, and asks in reverse to ensure pareto equality.

6. Bid/ask quantities can be partially settled.
7. A bid/ask quantity must be an integer multiple of 1 Wh. This is in consideration of future hardware integration, to allow direct use of the watt-pulse function of most smart-meters.
8. During the delivery period, if a seller is in short supply, it must financially compensate for the shortage at net metering prices. If batteries are available, the seller has the option to compensate by discharging its batteries, for all or part of the shortage during this period.
9. During the delivery period, if a buyer settled for more energy than used, the buyer must still pay the seller for the unused energy at the settlement price.

This market design strikes a compromise between a peer-to-peer market and a centralized market. By using pairwise settlements, a peer-to-peer like individualized value feedback can still be provided, while the simplicity and efficiency of a centralized market can be kept, especially for deployment in a small, localized region.

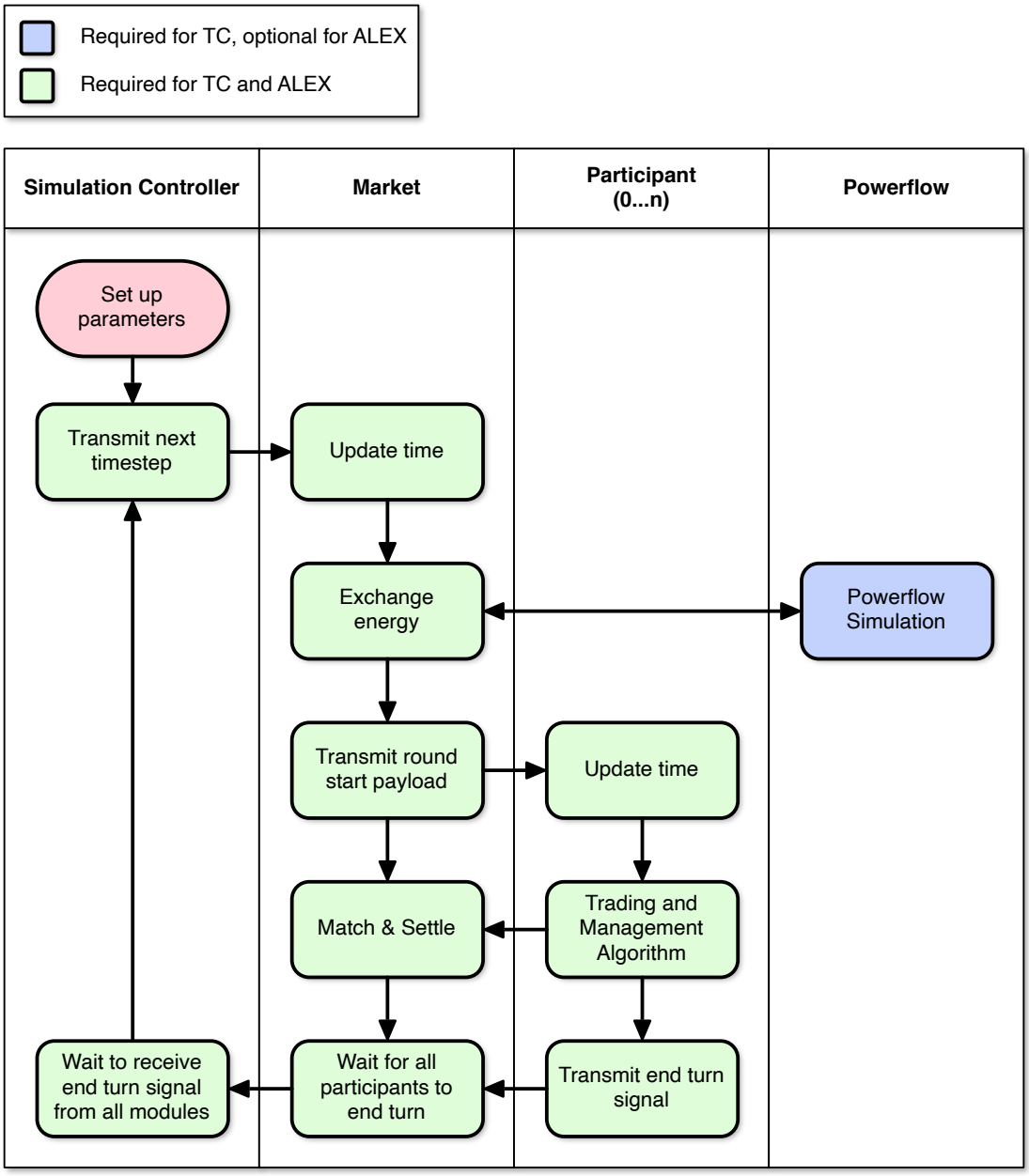

Figure B.13: Simplified swimlane diagram of TE schemes implemented in T-REX