SCALEX: SCALability EXploration of Multi-Agent Reinforcement Learning Agents in Grid-Interactive Efficient Buildings

Yara Almilaify, Kingsley Nweye and Zoltan Nagy
{yara.m,nweye,nagy}@utexas.edu
The University of Texas at Austin
Austin, TX, USA

ABSTRACT

Renewable energy transition and decarbonization pose significant challenges for grid-interactive efficient building communities. The optimization of intermittent renewable energy can be achieved using advanced control architecture and energy storage, enhancing energy flexibility. Reinforcement learning (RL) offers potential solutions, but its scalability and computational demands in large-scale settings remain unclear. This paper examines the scalability of Soft-Actor Critic (SAC) in multi-agent systems, comparing decentralized-independent SACs and centralized SACs using CityLearn, an OpenAI Gym environment. We consider neighborhoods consisting of 2 to 64 single-family residential buildings, each equipped with cooling and heating storage devices, domestic hot water storage devices, electrical storage devices, and solar PV systems. Our findings suggest that independent controllers outperform the centralized controller with increasing number of buildings. We also show that the performance on the building level can differ from the aggregated performance.

KEYWORDS

energy flexibility, demand response, multi agent system

ACM Reference Format:
Yara Almilaify, Kingsley Nweye and Zoltan Nagy. 2023. SCALEX: SCALability EXploration of Multi-Agent Reinforcement Learning Agents in Grid-Interactive Efficient Buildings. In *The 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '23), November 15–16, 2023, Istanbul, Turkey*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3600100.3623749

1 INTRODUCTION

Residential buildings contribute significantly to energy consumption and greenhouse gas emissions [5] making the transition to renewable energy sources and grid decarbonization critical for achieving zero-emission buildings [8]. However, these renewable sources can introduce grid instability due to supply-demand mismatches [18]. Grid-interactive efficient buildings (GEBs) can counteract these mismatches by providing flexibility services to the grid [10]. Although Demand Response (DR) programs have found success in commercial and industrial sectors, their application in residential buildings is still emerging and relies on distributed energy resources (DERs). Despite the potential of residential buildings to contribute to DR, their independent participation may not yield optimal performance at the aggregated level. Aggregators can enhance this performance by coordinating and selling energy flexibility. Advanced control algorithms, like Model Predictive Control (MPC) [3] and Reinforcement Learning Control (RLC) [12, 16, 17] have shown promise in various DER control applications. But their scalability, especially RLC's in multi-agent systems remains unclear.

In this paper, we aim to investigate this scalability issue. Specifically, we compare the performance of decentralized and centralized Soft-Actor Critic (SAC) agents: Using CityLearn, we assess the control performances in neighborhoods ranging from 2 to 64 single-family buildings. We seek to address the gap between SAC controllers in multi-agent systems and the scalability associated with managing residential energy flexibility. In particular, we will address the following questions: (1) *How does the performance of the controllers at different scales compare, and what are their advantages and limitations?* (2) *How do key performance indicators (KPIs) reflect the impact of scaling up the multi-agent system on load shaping and the management of residential energy flexibility?*

2 METHODOLOGY

2.1 Reinforcement Learning

Reinforcement learning (RL) is a machine learning subfield for sequential decision-making problems [13]. RL agents learn by interacting with the environment to maximize the rewards they receive. The Markov Decision Process (MDP) formalizes RL using a tuple (S, A, P, R), where agents interact using state (S) and action (A) spaces. At each timestep t, the agent observes a state (s_t), takes an action (a_t), and transitions to a new state (s_{t+1}) based on a transition probability function (P). The agent then receives a reward signal (r_{t+1}) using the reward function (R) to quantify its immediate performance. The objective of the agent is to learn the optimal control policy (π) that maximizes the expected cumulative reward.

Here, we are interested in cases where no prior information on the transition probabilities exists, and where both action and state space are continuous. Therefore, we explore the soft-actor critic (SAC) algorithm, a model-free, off-policy RL algorithm that has been employed in similar prior works [11]. The SAC architecture employs two deep neural networks to approximate the state-value function and the action-value function [6]. The actor network maps the current state to the action that it estimates to be optimal, while the critic network evaluates those actions by computing the value function [13]. SAC uses entropy maximization in order to maximize both entropy and expected rewards, which promotes greater policy

BuildSys '23, November 15–16, 2023, Istanbul, Turkey

ACM ISBN 979-8-4007-0230-3/23/11.
https://doi.org/10.1145/3600100.3623749

exploration, preventing the algorithm from prematurely converging to suboptimal solutions, and contributing to stable training. SAC also exhibits the advantage of reusing experiences, enabling efficient policy learning with fewer data samples, and handle complex and high-dimensional state and action spaces.

2.2 Simulation Environment

We use CityLearn, an OpenAI Gym environment designed to standardize multi-agent systems control [1, 15]. CityLearn provides a simulated environment of the demand response problem to flatten the aggregated curve of electrical demand by controlling energy storage devices of a variable number of buildings with diverse characteristics. Specifically, we analyze neighborhoods ranging from 2 to 64 buildings. These building models are created using physics-based energy models from the End-Use Load Profile (EULP) database and are representative of the single-family building stock in Austin, TX, USA [4]. The buildings have diverse load profiles and are equipped with electrical storage devices. The energy supply units are designed to satisfy the building's energy demand during the simulation by incorporating a backup controller that ensures that the energy supply units prioritize fulfilling the building's energy demands before storing any energy.

2.3 Action-State-Reward Design

We focus on residential battery energy storage system (BESS) control. The action space is a fraction of the battery capacity to be charged or discharged: the agent uses actions between -1 and 1 to discharge and charge the battery, respectively. The states are variables that can be observed within the environment but are not subject to direct control. They have a significant effect on the probability of receiving particular rewards. A total of 16 possible state variables can be categorized as temporal (calendar), weather, district, and building-specific states [1]. Detailed explanations of both action and state design can be found in [11]. For any agent i, the reward function is designed to minimize electricity consumption and maximize solar self-consumption to charge the BESS:

$$r_i^{SAC} = \sum_{i=0}^{n} -\left((1 + \frac{e}{|e|} \times storage_i^{SoC}) \times |e|\right) \quad (1)$$

where e is the energy use at time t (subscript omitted), and $storage_i^{SoC}$ is the state of charge of the BESS in building i. Thus, the reward function incentivizes achieving net-zero energy by imposing penalties when there is excess energy in BESS or when there's net export to the grid but the BESS is not fully charged. The penalty is maximized when the BESS is fully charged and there's a net import from the grid. However, there are no penalties or rewards when the BESS are fully charged during net exports. For centralized agents (see Section 2.4) the individual rewards in each building are aggregated into one reward function for the agent. For independent agents, there is no aggregation or reward exchange between the agents.

2.4 Control Architectures

We investigate two different types of control architectures: a fully decentralized scheme with independent agents, where no information is exchanged, and a fully centralized scheme where one agent controls the BESS in every building. A third, rule based controller (RBC) based on common practice is used as a reference.

2.4.1 Central SAC agents. $SAC_{central}$ enables a single agent to control all storage devices, facilitating information sharing of the state information of all buildings in the environment. Providing all information to one central agent presumably improves the performance of the agent at the aggregated neighborhood level, at the cost of longer computational and training times.

2.4.2 Decentralized-independent SAC agents. SAC_{indep} control the resources of their own building without knowing the other agents' policies and treat them as part of the environment. The agents independently learn and make decisions without explicit interaction or information sharing with other agents. We expect this to potentially result in better performance for individual buildings and deteriorating performance at the aggregated level.

2.4.3 Rule-Based-Control. We use an RBC as a baseline controller to compare the two SAC architectures. The RBC approach relies on a best-practice heuristic collection of if-then-else rules used to determine the best operating points within a control system [11]. However, these rules remain static, lacking the ability to automatically adapt to dynamic environmental change.

2.5 Key Performance Indicators (KPI)

We evaluate the controllers' performance using a set of KPIs that are to be minimized: electricity consumption and zero net energy on individual building level, and (annual) peak demand, average daily peak, ramping, and 1-Load factor on the aggregated district level (see [11] for details). The KPIs account for the deterministic action taken after the reward has reached convergence. KPIs of the RL agents that exceed the corresponding RBC indicate poorer performance, while lower values indicate superior performance.

3 RESULTS

We perform our simulations on Stampede2, a high-performing computer at the Texas Advanced Computing Center [2]. Stampede2 hosts 4,200 Knights Landing (KNL) nodes, each equipped with a single processor at 1.4 GHz and 96 GB RAM. Parallelization capabilities were also utilized to optimize the training process. Understanding these constraints is important as they directly impact the total training time and the feasibility of implementing the study. The simulations consist of 50 episodes, each spanning 8760 time steps with each time step equivalent to one hour, thereby representing the hours in a year. We decided on 50 trials by trial and error, and also due to our prior extensive experience working with CityLearn. To ensure robustness, we repeat each simulation three times with random seeds and averaging the results. We assume no distribution shift between the training and the deployment of the RL agent.

Figure 1 shows the training curve of both SAC algorithms. Both reward curves are normalized to the last reward value which is used to scale the rewards between 0 and 1, enhancing interpretability and facilitating a meaningful comparative analysis of their performance. Initially, the agent explores, resulting in higher rewards. This indicates that the agent is learning to perform better. As it learns, it converges towards optimal policies, indicating improved performance and stable strategy exploitation. As the number of buildings

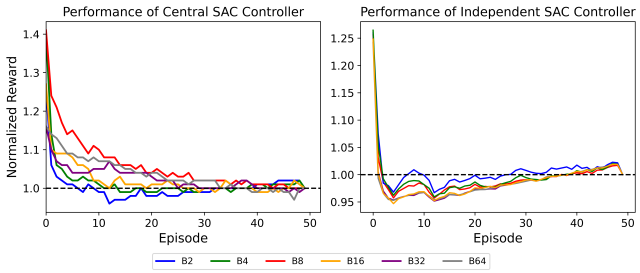

Figure 1: Reward curve of central SAC (left) and independent SAC (right) agents, normalized to the final reward value. At episode 50, the agents use the deterministic action.

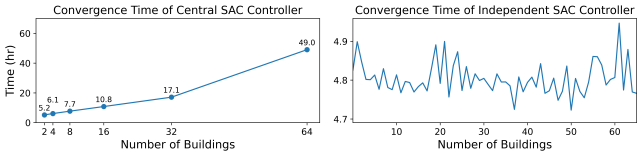

Figure 2: Convergence time for central (left) and independent (right) agents.

varies from 2 to 64, SAC_{indep} demonstrates faster convergence by leveraging its independence and smaller state space. In contrast, $SAC_{central}$ exhibit slower convergence due to the complexities of controlling multiple resources, requiring more training time/data to converge to an optimal policy. Overall, both curves eventually converge, indicating that we have simulated a sufficient amount of time to observe steady-state behavior. It's crucial to recognize that this convergence does not imply that both agents perform similarly as the agent-environment setup is different.

Figure 2 shows the impact of scaling up the environment on the simulation time required to reach convergence. We assume the SAC_{indep} approaches convergence at about 30 episodes, taking on average about 4.8hr. Similarly, $SAC_{central}$ approaches convergence at around 30 episodes, but the convergence time gradually increases from 5.2 to 49hr. This aligns with our earlier hypothesis, suggesting that the complexity and scale of the environment impact the convergence dynamics. Specifically, training larger districts under centralized SAC controllers become computationally intensive, resource-demanding, and costly.

Figure 3 compares the building-level KPIs, electricity consumption and zero net energy, for varying the number of buildings. Clearly, the SAC_{indep} outperforms both $SAC_{central}$ and RBC, resulting in an average reduction of 33% in electricity consumption compared to the RBC. For very few buildings (2-4), the performance of $SAC_{central}$ is similar to SAC_{indep} (30% avg reduction). However, as the number of buildings increases, the performance of $SAC_{central}$ gradually declines (9% avg reduction) relative to the SAC_{indep}, though still outperforming the RBC. For a larger number of buildings (32-64), we find that the performance for some buildings controlled with $SAC_{central}$ is similar to the RBC, potentially negating all advantages of advanced controllers (for this KPI). The difference between the approaches lies in the agents' control approach, where SAC_{indep} effectively leverages their autonomy to explore the environment and allocate resources to their own

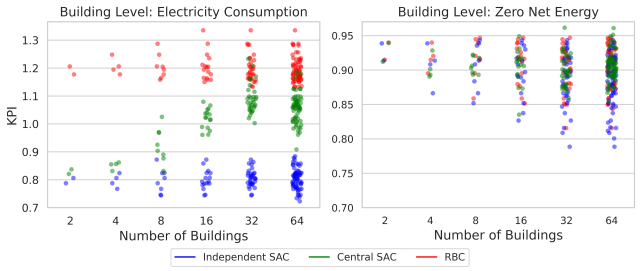

Figure 3: Building-level KPI for each control architecture and varying the number of buildings. Each dot is one building.

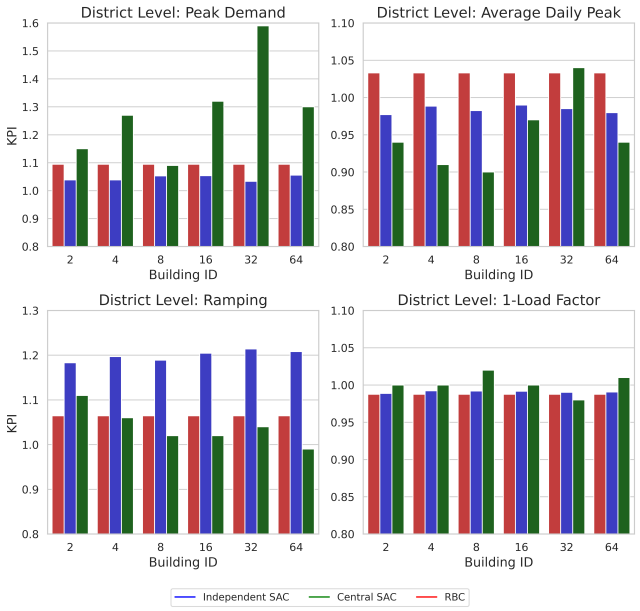

Figure 4: District-level KPI for each control architecture and varying the number of buildings. The y-axis is exaggerated, starting at 0.8, to emphasize the differences in KPI values.

building, leading to a greater overall reduction at the building-level. Additionally, both SAC algorithms show no significant impact on the zero net energy metric.

Figure 4 compares the (aggregated) district-level KPIs. The SAC_{indep} outperforms the RBC slightly by about 9% on the annual peak demand. In contrast, the $SAC_{central}$ does not exhibit peak demand reduction ability as it gradually performs worse as the number of buildings increases. Both the $SAC_{central}$ and SAC_{indep} effectively reduce the average daily peak, showcasing their capacity to shape the daily load curve. It is somewhat unclear why $SAC_{central}$ seems to display no correlation with the number of buildings. We speculate that it could be due to the fact that the peak demand KPI is a single value measured over more than 8000 timesteps, so the controller cannot average out a one time poor performance.

In Figure 4, we also observe that SAC_{indep} exhibits poorer performance for ramping compared to the RBC, while $SAC_{central}$ exhibits some adaptability to varying demand and supply conditions. As for the 1-Load factor KPI, both SAC algorithms perform poorly

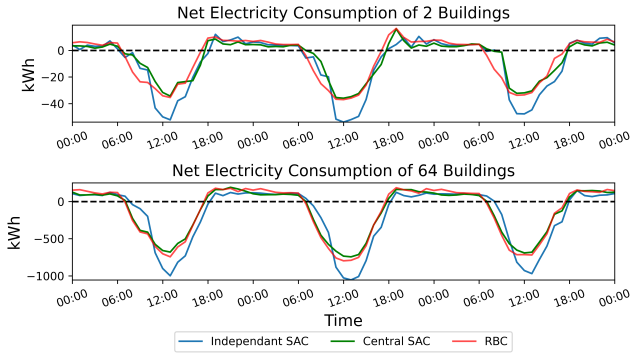

Figure 5: Comparing the net load profile for 2 buildings (top) and 64 buildings (bottom) during three days in July.

compared to RBC, indicating that controllers did not efficiently utilize the building's available capacity.

Figure 5 shows the hourly district-level net load profile for the first 3 days of July, during which the demands are higher and load shaping is more critical. Between the hours of 19:00 and 00:00 for 64 buildings, the central and independent SAC agents reduced the energy usage by an average of 8.88% and 55.08%, respectively, compared to the RBC. During periods with no sunlight, the SAC agents effectively offset the load by utilizing the stored battery charge during higher peak periods. Despite the slight drop, the independent SAC demonstrates the best performance in load shaping compared to both the central SAC and RBC. As expected, the net load of the larger group (64) is smoother compared to 2 buildings due to averaging.

4 DISCUSSION AND CONCLUSION

Our research shed light on the SAC controllers' performance for an increasing number of buildings in the environment. We found that $SAC_{central}$ has slower convergence and requires more training time. The advantage of SAC_{indep} lie in their faster convergence due to their independence, enabling them to explore their environment more efficiently at lower computational cost.

We find that scaling up the multi-agent system has notable impact on load shaping and residential energy flexibility management. The SAC_{indep} agent was able to reduce most of the KPI metrics but struggled to reduce ramping and 1-load factor. The $SAC_{central}$ also reduced most of the KPI metrics, except for peak demand and the 1-load factor. This may be influenced by the agents' centralized approach when it receives information on all buildings, potentially allocating excessive energy for certain buildings. On the other hand, since SAC_{indep} autonomously learns and adapts to its own policies, the collective individual optimization across multiple buildings may not always lead to the best overall performance of the entire system.

Our work demonstrates that it is critical to evaluate and integrate multiple KPIs when optimizing grid-interactive buildings. Most research focuses on one or few. However, they are interdependent to achieve energy flexible buildings. For instance, the annual peak load KPI is related to infrastructure planning, e.g. transmission and distribution networks, while the average daily demand KPI is relevant for reducing the need for peaker plants.

Recent research has shown that building load coordination control is important to improve grid reliability and reduce infrastructure cost [9, 14]. Our work suggests that centralized RL approaches are more challenging to achieve while offering worse performance compared to decentralized solutions. Similar findings have been proposed for decentralized model-predictive control [7].

Future research in this area should focus on exploring potential areas of improvement in the design of the reward function to further enhance load shifting and shaping capabilities, as well as improve the performance of the controllers. Another idea is to implement cooperative multi-agent SAC frameworks instead of a centralized controller as it may be possible to reduce the action space for the controller, leading to sufficient exploration.

REFERENCES

[1] [n. d.]. https://github.com/intelligent-environments-lab/CityLearn
[2] [n. d.]. TACC - Stampede2 User Guide Documentation. https://docs.tacc.utexas.edu/hpc/stampede2/
[3] Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L. Vrabie, and Lieve Helsen. 2020. All you need to know about model predictive control for buildings. , 190–232 pages.
[4] Wilson et al. 2022. End-Use Load Profiles for the U.S. Building Stock: Methodology and Results of Model Calibration, Validation, and Uncertainty Quantification. (2022). https://doi.org/10.2172/1854582
[5] Benjamin Goldstein, Dimitrios Gounaridis, and Joshua Newell. 2020. The carbon footprint of household energy use in the United States. *Proceedings of the National Academy of Sciences* 117 (7 2020), 201922205.
[6] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. (1 2018). http://arxiv.org/abs/1801.01290
[7] Nicolas Lefebure, Mohammad Khosravi, Mathias Hudobade Badyn, Felix Bünning, John Lygeros, Colin Jones, and Roy S. Smith. 2022. Distributed model predictive control of buildings and energy hubs. *Energy and Buildings* 259 (3 2022).
[8] Benjamin D. Leibowicz, Christopher M. Lanham, Max T. Brozynski, José R. Vázquez-Canteli, Nicolás Castillo Castejón, and Zoltan Nagy. 2018. Optimal decarbonization pathways for urban residential building energy services. *Applied Energy* 230 (11 2018), 1311–1325. https://doi.org/10.1016/j.apenergy.2018.09.046
[9] Thomas Navidi, Abbas El Gamal, and Ram Rajagopal. 2023. Coordinating distributed energy resources for reliability can significantly reduce future distribution grid upgrades and peak load. *Joule* (2023).
[10] Monica Neukomm, Valerie Nubbe, and Robert Fares. 2019. *Grid-Interactive Efficient Buildings Technical Report Series: Overview of Research Challenges and Gaps.* Technical Report. United States.
[11] Kingsley Nweye, Siva Sankaranarayanan, and Zoltan Nagy. 2023. MERLIN: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities. (2023). https://doi.org/10.18738/T8/0YL
[12] Daniel O'Neill, Marco Levorato, Andrea Goldsmith, and Urbashi Mitra. 2010. Residential Demand Response Using Reinforcement Learning. In *2010 First IEEE International Conference on Smart Grid Communications.* 409–414.
[13] Richard Sutton and Andrew Barto. 2018. Reinforcement Learning: An Introduction (Second Edition). (2018).
[14] José Vázquez-Canteli, Gregor Henze, and Zoltán Nagy. 2020. MARLISA: Multi-Agent Reinforcement Learning with Iterative Sequential Action Selection for Load Shaping of Grid-Interactive Connected Buildings. *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation.*
[15] José R Vázquez-Canteli, Sourav Dey, Gregor Henze, and Zoltán Nagy. 2020. CityLearn: Standardizing Research in Multi-Agent Reinforcement Learning for Demand Response and Urban Energy Management. (2020). https://arxiv.org/abs/2012.10504
[16] Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (7 2020).
[17] Lei Yang, Zoltan Nagy, Philippe Goffin, and Arno Schlueter. 2015. Reinforcement learning for optimal control of low exergy buildings. *Applied Energy* 156 (2015), 577–586. https://doi.org/10.1016/j.apenergy.2015.07.050
[18] Mohammed Yekini Suberu, Mohd Wazir Mustafa, and Nouruddeen Bashir. 2014. Energy storage systems for renewable energy power sector integration and mitigation of intermittency. , 499–514 pages.