Combining Machine Learning and Control Theory to Enhance the Operation of Energy Systems

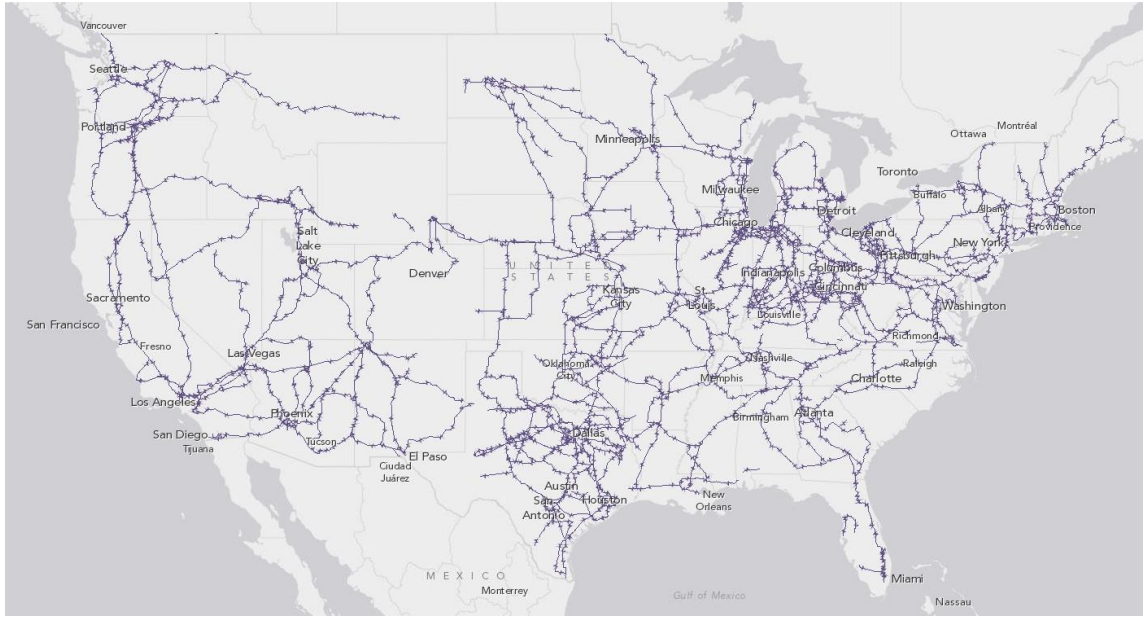
ACM eEnergy 2024

Omid Ardakanian
Associate Professor
Department of Computing Science

UNIVERSITY OF ALBERTA

How to reliably control energy systems?

Scale and complexity

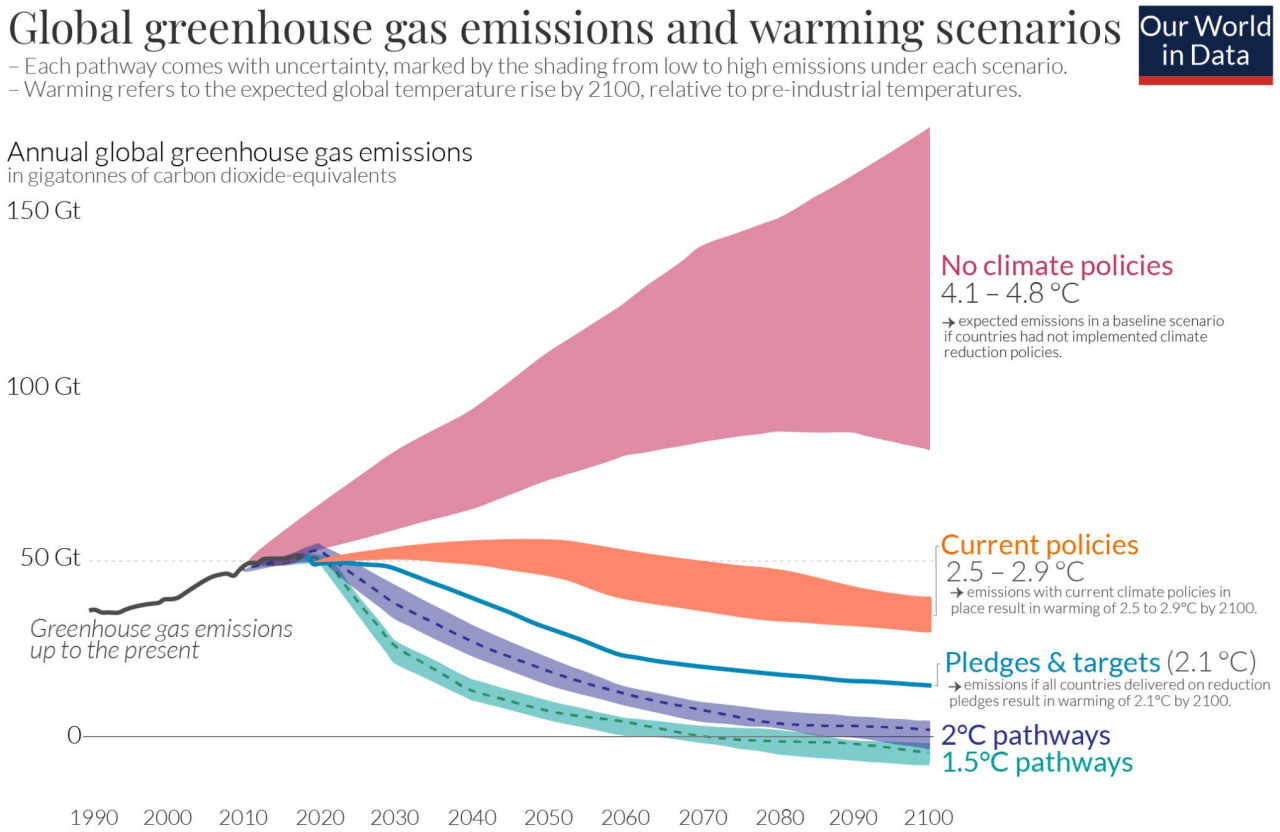


nonlinear dynamics, **safety, stability and worst-case performance requirements**, numerous components and interfaces, adversarial and strategic decisions by humans and control agents

But they are far from optimized!

- reliability, safety and stability come before optimality
- high variability and uncertainty
- partial observability
  - e.g. only a few sensors installed beyond the substation, zone occupancy is not measured directly, …
- human-in-the-loop issues
  - e.g. setpoints and occupancy schedules defined by facilities managers

Urgent need to reduce emissions of future energy systems



Urgent need to learn and incorporate time-varying dynamics and constraints into the operation of future energy systems

Satellite image taken during Winter Storm Uri; Source: National Oceanic and Atmospheric Administration (NOAA)

Designing future energy systems

Requirements

Functional requirements

- increase efficiency, adaptability, and autonomy
- ensure safe and stable operation at design time and after deployment
  - 3 safety levels: violations are discouraged, no violations with high probability, no violations
- ensure robustness against adversarial and strategic behaviors

Non-functional requirements

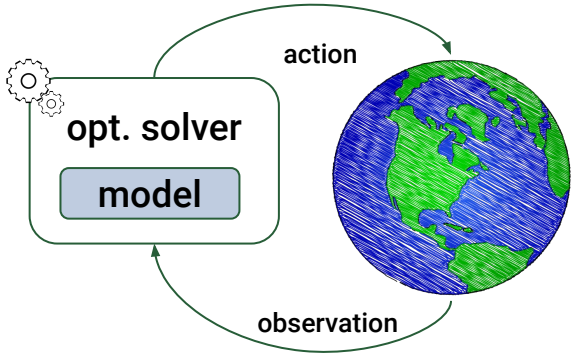
- controls must be simple and easy to implement
- operators and domain experts must find controls intuitive

Designing future energy systems

Principles

- build off of existing controllers when possible
  - there is already a stabilizing controller for most energy systems
- guarantee safety in probability or at all times
  - discouraging constraint violation is not enough
- pay attention to performance in worst-case scenarios
  - focusing on average-case performance only can be disastrous
- design data-efficient learning algorithms
  - "useful” data is not abundant

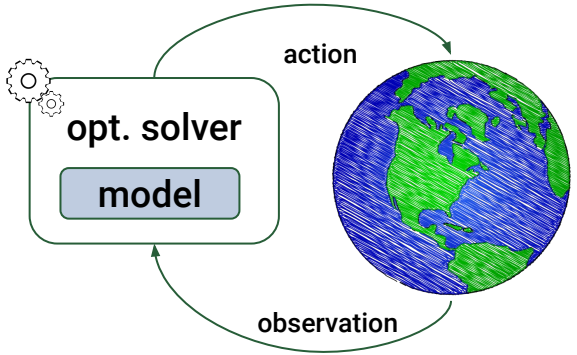
Control-theoretic approaches



Benefits:

- optimize performance given a prior system dynamics model
- constraints (safety, stability, risk) can be handled
- theoretical guarantees for stability and performance under (bounded) uncertainty

Control-theoretic approaches



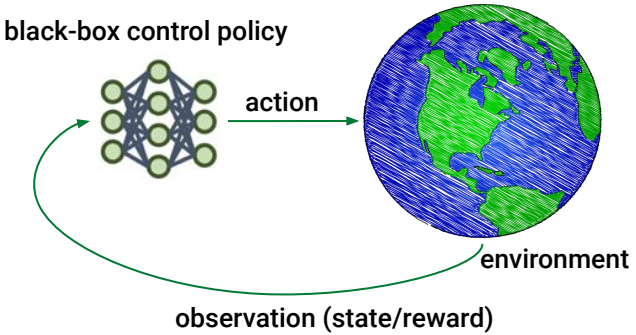
Benefits:

- optimize performance given a prior system dynamics model
- constraints (safety, stability, risk) can be handled
- theoretical guarantees for stability and performance under (bounded) uncertainty

Challenges:

- achieving high performance requires a sufficiently accurate model
- high computation cost
  - limited to specific families of cost functions and constraints
  - need to opt for a short prediction horizon, e.g. when there are many control variables
- lack of generalization to new contexts

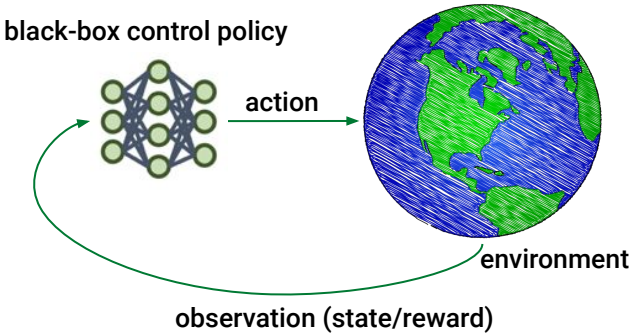
Learning-based approaches



Benefits:

- optimal control is learned* from offline or online data
  - prior or learned model can help but is not necessary
- low-cost implementation once training completes
- endless adaptation to changing environment with continual learning

Learning-based approaches



Benefits:

- optimal control is learned* from offline or online data
  - prior or learned model can help but is not necessary
- low-cost implementation once training completes
- endless adaptation to changing environment with continual learning

Challenges:

- neural networks may bend the laws of physics and violate constraints!
- opaque decision making
- relationship between data need and achievable performance is not well understood
- convergence is not guaranteed when neural networks are used to approximate value function or policy
- learning from interaction with the real environment is potentially risky

Can we bring together the best of each approach?

Extensive work in this area

- integrating learning with model predictive control (MPC)
  - learning dynamics and cost via
    end-to-end learning with imitation loss or reinforcement learning reward,
    convex body chasing,
    black-box optimization, …
  - learning terminal cost and constraint (e.g. using a neural network that approximates value function)
- PID controller tuning using reinforcement learning
- learning to switch between high-performance and high-assurance controllers
- …

Extensive work in this area

- incorporating physics and constraints into a learning-based controller **during training**
  - physics-informed/corrected neural network
  - implicit optimization layer for constraint satisfaction in neural network
  - differentiable MPC as policy instead of a generic neural network
  - risk-averse and constrained reinforcement learning
  - …
- imposing constraints on a learning-based controller **after training**
  - using model-based advice, e.g. model predictive safety/stability filter
  - through combination with one or multiple model-based controllers
  - …

My work in this area

- learning system dynamics — TCNS'19, ENB'21, TCNS'23
- learning a convex cost function for optimization and control — eEnergy'24
- learning to tune a feedback controller — eEnergy'20, TSG'21
- learning constrained RL policies — eEnergy'23, BuildSys'22
- learning a finite pool of RL policies and choosing the most suitable one for transfer to a novel environment — ENB'23

My work in this area

- learning system dynamics — TCNS'19, ENB'21, TCNS'23
- learning a convex cost function for optimization and control — eEnergy'24
- learning to tune a feedback controller — eEnergy'20, TSG'21
- learning constrained RL policies — eEnergy'23, BuildSys'22
- learning a finite pool of RL policies and choosing the most suitable one for transfer to a novel environment — ENB'23

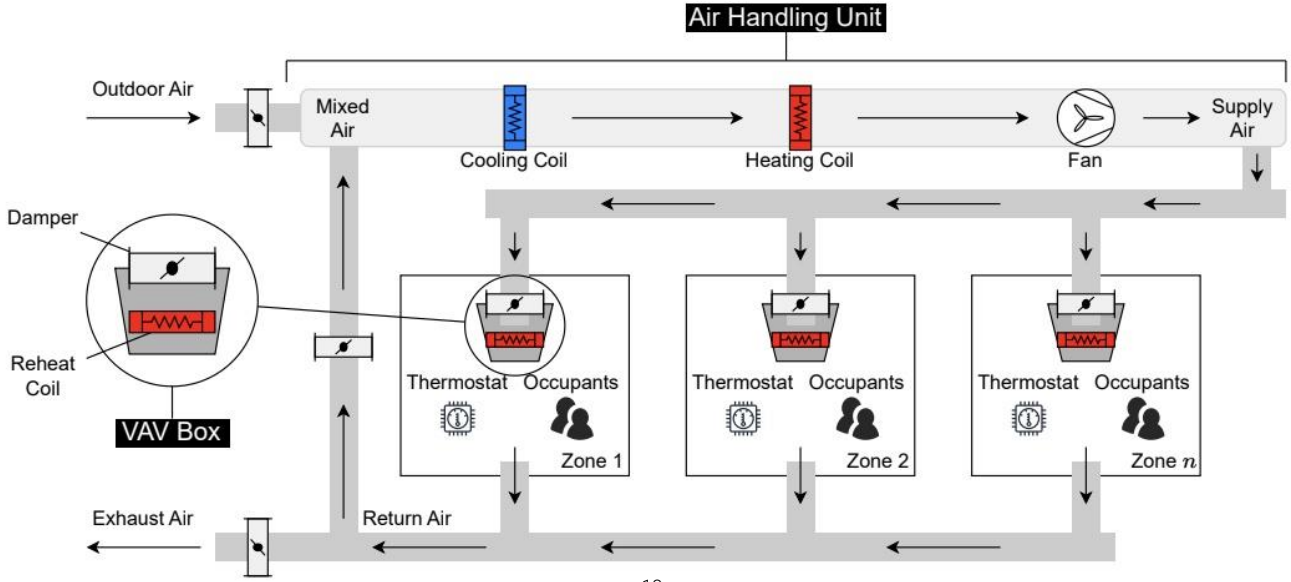


Learning a pool of diverse black-box policies and transferring the most suitable one

Problem statement

minimize HVAC energy consumption while keeping the temperature of occupied zones within bounds

- decoupling AHU and VAV control problems is common



Control schemes

- Existing controller
  - satisfies thermal comfort requirements by maintaining zone temperature around its setpoint
  - performance is not optimal with respect to energy consumption
- MPC
  - a sufficiently accurate model must be developed for each building which is labour intensive and costly
- RL policy
  - ensuring no constraint violation (e.g. thermal comfort) is challenging
  - convergence to asymptotic performance is slow when state/action space is large

Control schemes

- Existing controller
  - satisfies thermal comfort requirements by maintaining zone temperature around its setpoint
  - performance is not optimal with respect to energy consumption
- MPC
  - a sufficiently accurate model must be developed for each building which is labour intensive and costly
- RL policy
  - ensuring no constraint violation (e.g. thermal comfort) is challenging
  - convergence to asymptotic performance is slow when state/action space is large
- **Our approach (RL policy + existing controller)**
  - adopt RL policy to continuously adjust a few points (temperature setpoint or minimum damper position) and keep using the existing controller to control other knobs in AHU and VAV

Control policy for HVAC

Background on deep reinforcement learning

1. Generate samples

Value-based

2. Fit Q(s,a)

3. $\pi(s) \leftarrow \text{argmax}_a\ Q(s_t,a_t)$

Policy-based

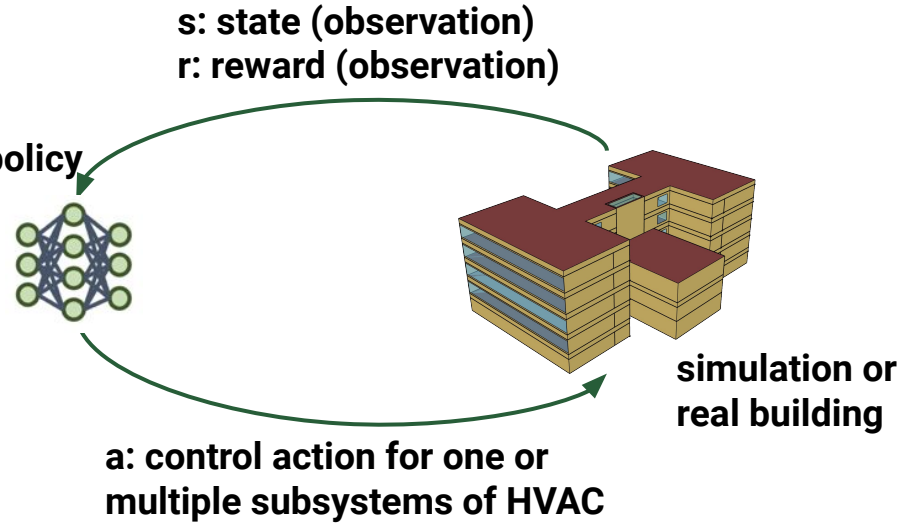
2. Evaluate $R = \Sigma_t\ r(s_t,a_t)$

3. $\theta \leftarrow \theta + \alpha \nabla_\theta \mathbf{E}_\pi\ [\Sigma_t\ r(s_t,a_t)]$

Actor-Critic

2. Fit V(s) or Q(s,a) to sampled reward sum

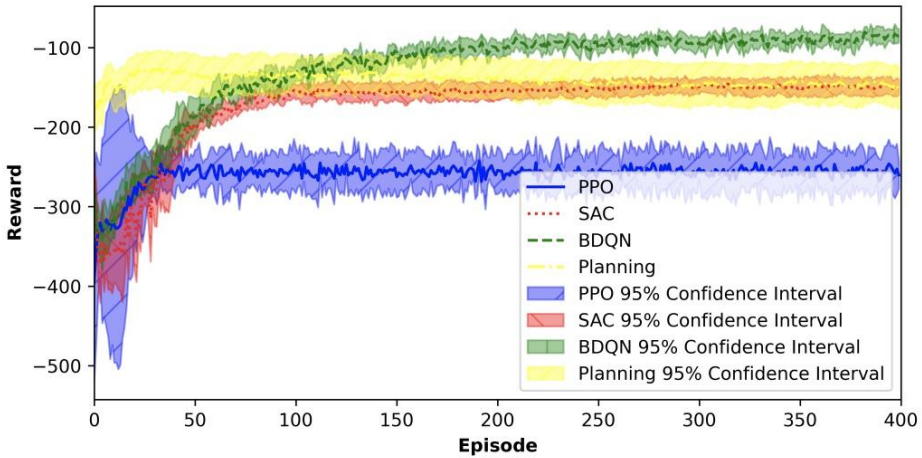
3. $\theta \leftarrow \theta + \alpha \nabla_\theta \mathbf{E}_\pi\ [Q(s_t,a_t)]$



Learning is quite slow in the target environment

it may take **several months** (or **years)** to learn a high-quality control policy

result obtained in a 5-zone building simulated in EnergyPlus & COBS



Tianyu Zhang, Gaby Baasch, Omid Ardakanian, Ralph Evins, "On the Joint Control of Multiple Building Systems with Reinforcement Learning", In Proceedings of the 12th ACM International Conference on Future Energy Systems (ACM e-Energy), pp. 60-72, 2021.

Practical challenges

- if agent interacts with the real building, poor performance is likely in the early stage of training
- if agent interacts with a simulator, discrepancies between real and simulated environments could lead to poor performance after deployment

Practical challenges

- if agent interacts with the real building, poor performance is likely in the early stage of training
- if agent interacts with a simulator, discrepancies between real and simulated environments could lead to poor performance after deployment

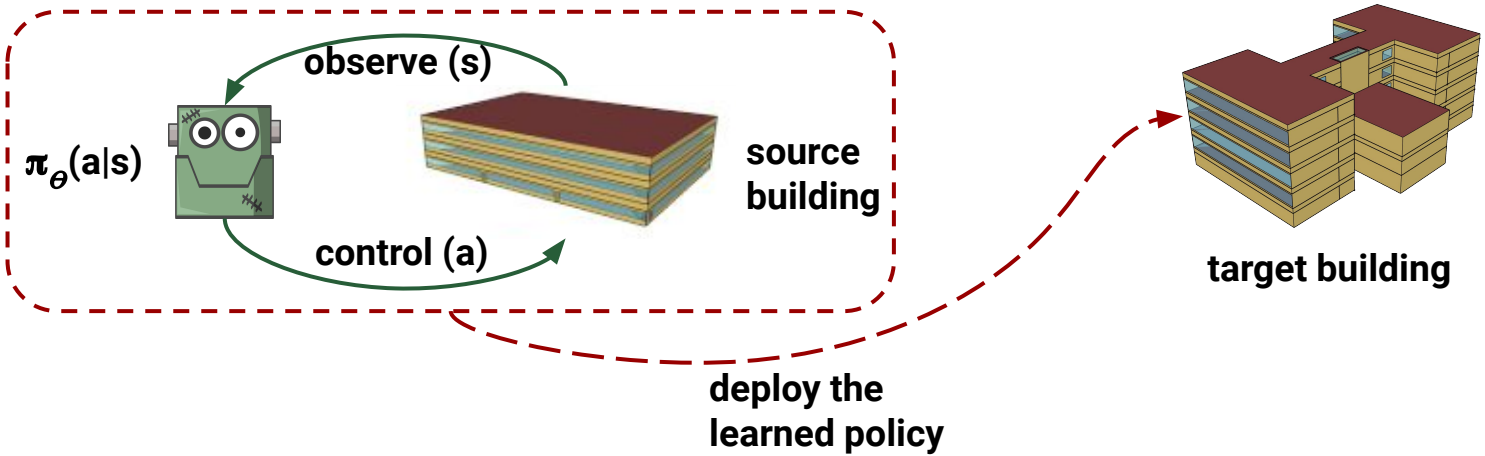
Question: how to reduce the training cost of RL so we can learn a high-quality policy without a high-fidelity simulator?

How to reduce training cost of RL agents?

Transfer learning approach

train RL agent through interaction with a **reference commercial building** (e.g. a test facility), deploy to the target building, and retrain

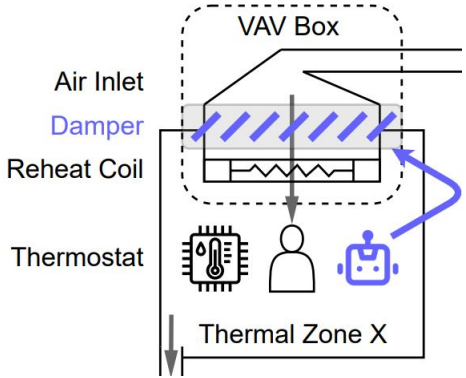
Problem: source and target buildings may have a different amount of sense and control points so we have different MDPs!



How to reduce training cost of RL agents?

Source and target buildings may have different control knobs

use the **multi-agent reinforcement learning** framework in which each agent is responsible for controlling a single zone

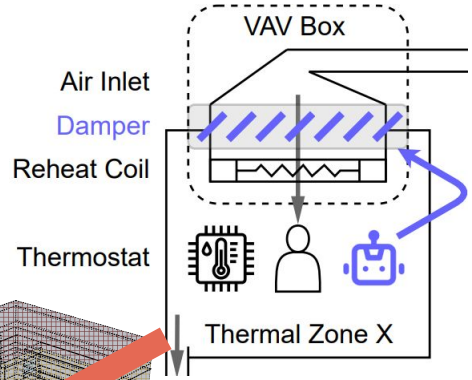


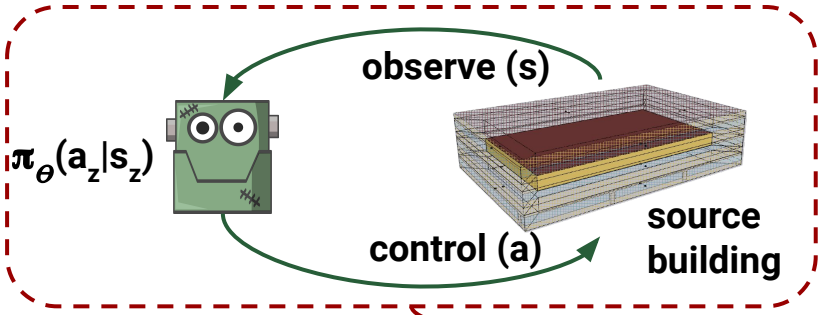
How to reduce training cost of RL agents?

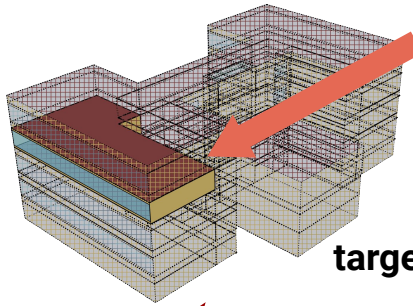
Source and target buildings may have different control knobs

use the **multi-agent reinforcement learning** framework in which each agent is responsible for controlling a single zone

Problem: learned policy in the source building may still perform poorly in the target building because they have different dynamics







target building

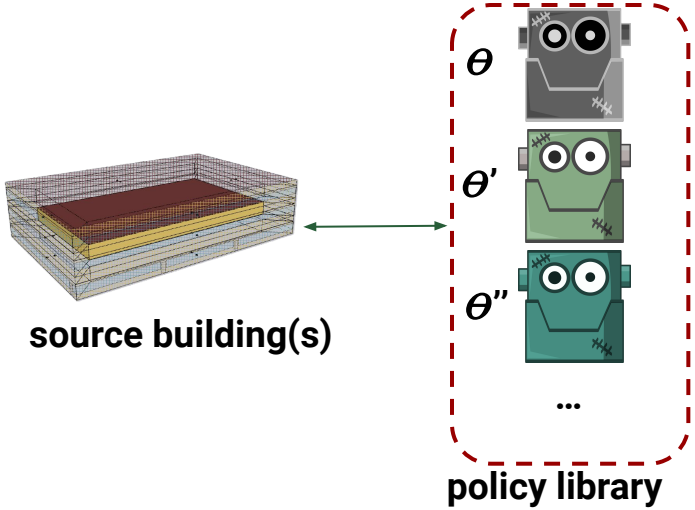
deploy the learned policy to the corresponding zone

How to reduce training cost of RL agents?

Choosing a learned policy from a finite pool of candidates

train a **finite pool** of RL agents by exposing them to different dynamics, then identify and assign the **most suitable** one to each zone of the target building

Problem: a single environment may not exhibit (many) distinct dynamics

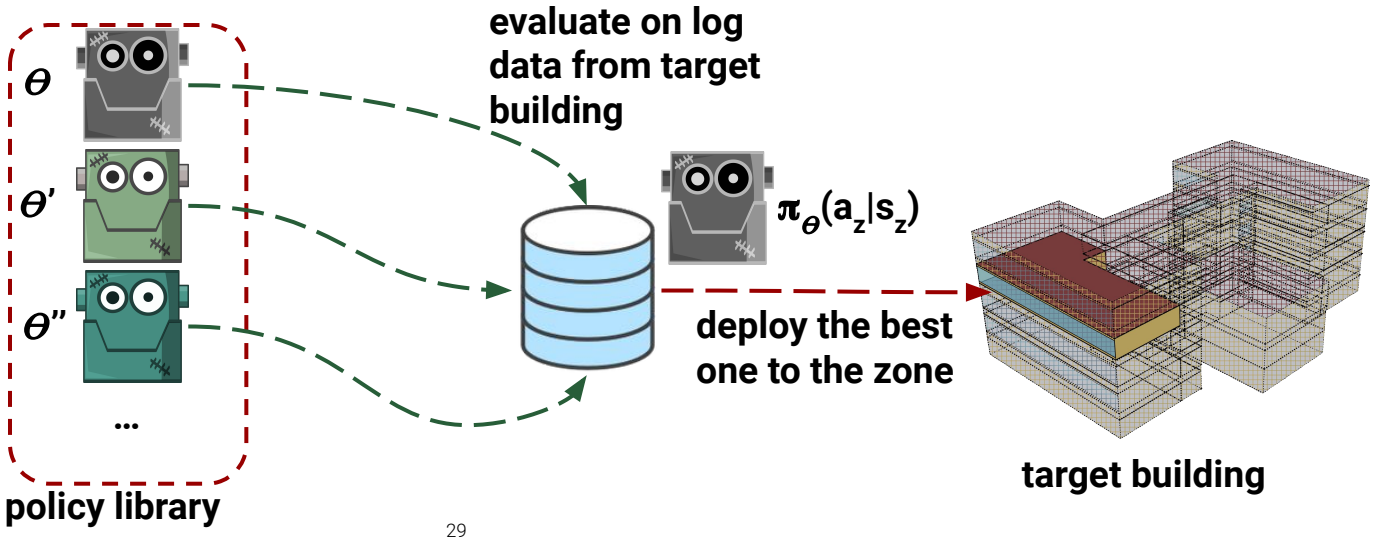


How to reduce training cost of RL agents?

Choosing a learned policy from a finite pool of candidates

train a **finite pool** of RL agents by exposing them to different dynamics, then identify and assign the **most suitable** one to each zone of the target building

Problem: a single environment may not exhibit (many) distinct dynamics



Policy Diversity and Evaluation

- there are not many environments (test facilities) that can be used for training, so we learn **diverse policies** in each to get a larger pool of candidate controllers

- then we **efficiently** identify the best policy in the pool for transfer to each zone

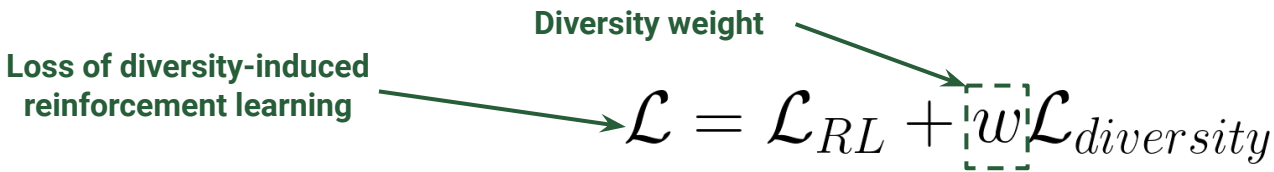
Policy Diversity and Evaluation

- there are not many environments (test facilities) that can be used for training, so we learn **diverse policies** in each to get a larger pool of candidate controllers
- then we **efficiently** identify the best policy in the pool for transfer to each zone

Intuition: **the specific skill learned for one task may not be useful in another, but having a diverse set of skills could be helpful in a novel task**

Different kinds of diversity

- augmenting the loss function of a policy gradient RL algorithm (**policy diversity**)
  - the policy must be different from previously learned policy in addition to maximizing the expected return



Aakash Krishna, Tianyu Zhang, Omid Ardakanian, Matthew Taylor, "Mitigating an Adoption Barrier of Reinforcement Learning-Based Control Strategies in Buildings", Energy and Buildings 285 (2023): 112878.

Diversity loss

Set of previously learned policies

$$\mathcal{L}_{diversity} = -\frac{\sum_{\pi' \in \Pi_{learned}} \sum_{(s,a) \in \exp} \frac{max\left(\frac{\max\left(\pi(a|s), \pi'(a|s)\right)}{\min\left(\pi(a|s), \pi'(a|s)\right)}, \bar{\rho}\right)}{\left|G^{\exp}(s) - V^{\pi'}(s)\right|}}{|\Pi_{learned}|}$$

Aakash Krishna, Tianyu Zhang, Omid Ardakanian, Matthew Taylor, "Mitigating an Adoption Barrier of Reinforcement Learning-Based Control Strategies in Buildings", Energy and Buildings 285 (2023): 112878.

Different kinds of diversity

- augmenting the loss function of a policy gradient RL algorithm (**policy diversity**)
  - the policy must be different from previously learned policy in addition to maximizing the expected return

- learning policies on multiple environments (**environmental diversity**)
  - the more environments we see during training, the lower would be our uncertainty

Aakash Krishna, Tianyu Zhang, Omid Ardakanian, Matthew Taylor, "Mitigating an Adoption Barrier of Reinforcement Learning-Based Control Strategies in Buildings", Energy and Buildings 285 (2023): 112878.

Policy evaluation/selection methods

- **brute force:** run every policy in the library on each zone of the target building, use energy consumption under this policy as evaluation metric, and transfer the best policy determined in this way to that zone
  - extremely expensive
  - yields a lower bound on HVAC energy consumption through policy transfer

Policy evaluation/selection methods

- **off-policy evaluation** (OPE): evaluate the performance of policies learned in a training environment without online interaction with the target environment
  - IPW and SNIPW use importance and rejection sampling to reweight the rewards according to **log data** obtained from the target environment

    Gaussian Kernel (GK) uses kernel density estimation instead of rejection sampling for continuous actions

Policy evaluation/selection methods

- **zero-cost proxy** (ZCP) is used in the domain of neural architecture search to rank different architectures at initialization
  - gradnorm (GN) and single-shot network pruning (SNIP) compute the loss and its gradient for a minibatch of data (**log data** obtained from the target environment)
  - since our control policies are neural networks, ZCPs can be used to rank them!

MARL framework

State as perceived by each agent

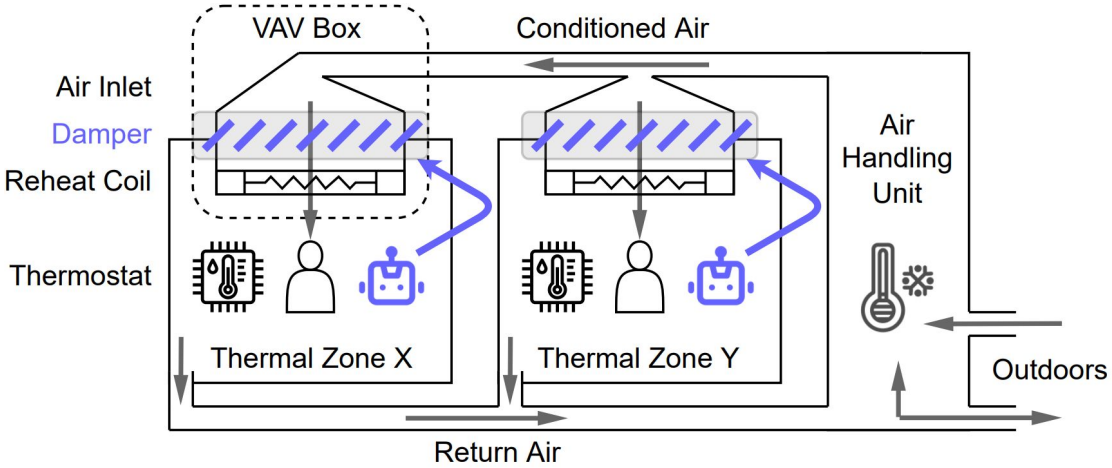
- zone temp.
- zone humidity
- zone occupancy
- outdoor temp.
- solar radiation
- hour of the day

Action of each agent

- minimum damper position

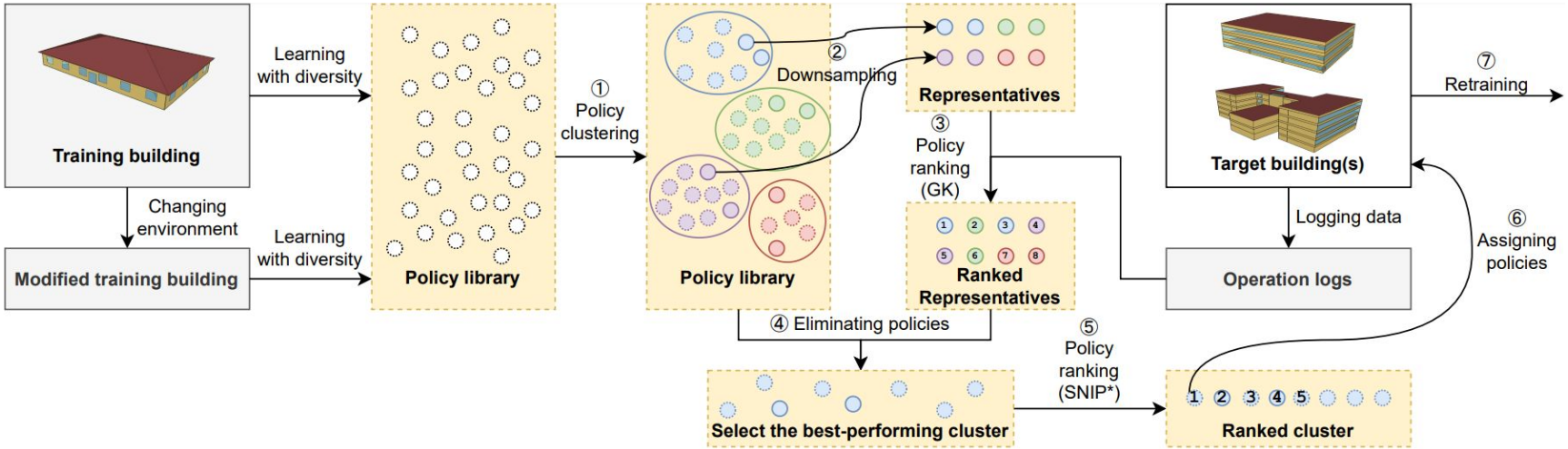
Reward provided to each agent

- energy use of the respective VAV system



The big picture

all policies learned using PPO

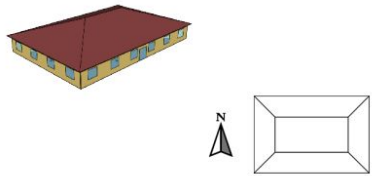


Source and target buildings

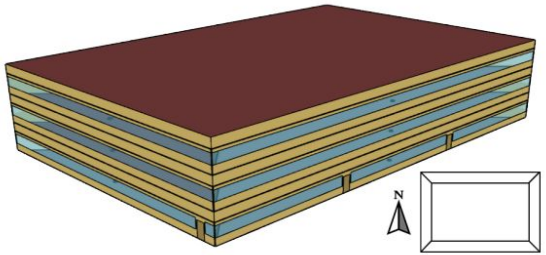
- Building A: 1-story 5-zone small office (511.16 m^2), located in Denver, US
- Building B: 3-story 15-zone medium office (4,982.19 m^2)
- Building C: 5-story 26-zone real building (5,051 m^2), located in San Francisco, US

all experiments were run in COBS*

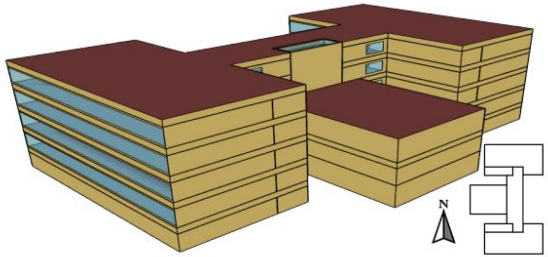
Training



(a) Building A



(b) Building B_{Denver}



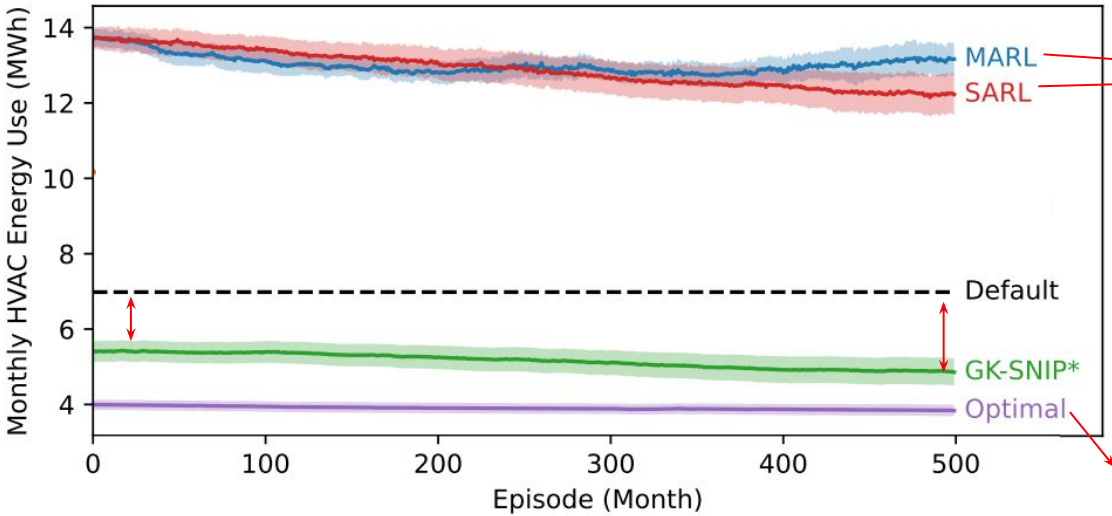
(c) Building C

* Tianyu Zhang, Omid Ardakanian, "Poster Abstract: COBS: COmprehensive Building Simulator",
In Proceedings of the 7th ACM Conference on Systems for Built Environments (BuildSys), November 2020.

Experiment result

Transfer to Building B_{Denver} (same city)

Policy evaluation/selection is done using 15 days of log data from Building B_{Denver}



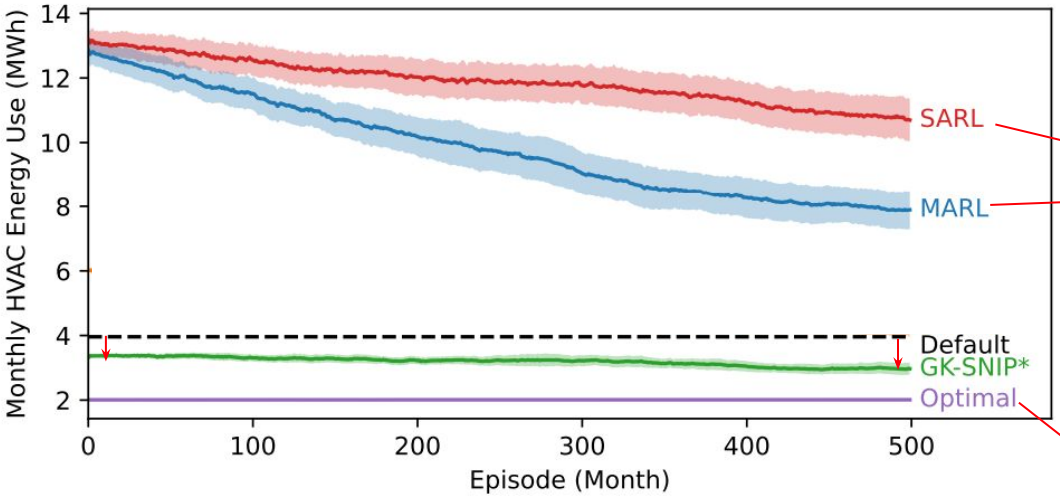
agent(s) trained on target building only (no knowledge transfer)

if we knew the best policy assignment to zones in the target building

Experiment result

Transfer to Building $B_{SanFrancisco}$

Policy evaluation/selection is done using 15 days of log data from Building $B_{SanFrancisco}$



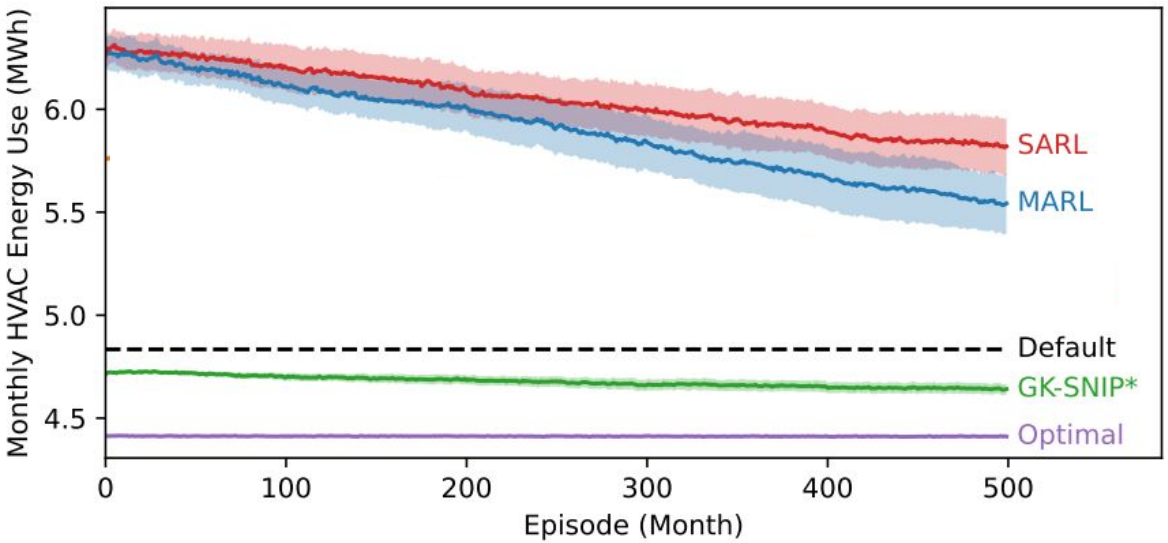
agent(s) trained on target building only (no knowledge transfer)

if we knew the best policy assignment in the target building (lower bound)

Experiment result

Transfer to a real building (Building C)

Policy evaluation/selection is done using 15 days of log data from Building C



Conclusion

- **up to 30.4%** energy savings can be achieved early on
- proposed controller outperforms default rule-based controller **even before retraining** in the target building
- can we use a combination of the top K policies rather than picking the top-ranked policy to control a zone?
    - tune weights using an online algorithm…

Summary

- combining machine learning and control techniques is a promising direction
  - higher control performance with formal guarantees for safety and stability
- dealing with abruptly changing and slowing drifting dynamics is still challenging
- reinforcement learning has been successful in discovering drugs and game playing strategies, but “discovering” black-box control policies will not make the cut in safety-critical systems
  - design-time guarantees, efficient run-time recovery mechanisms, and explainability are required to build **trust**

ardakanian@ualberta.ca
https://cs.ualberta.ca/~oardakan/