Peak shaving in district heating exploiting reinforcement learning and agent-based modelling

Peak Shaving in District Heating exploiting Reinforcement Learning and Agent-Based Modelling

Francesco M. Solinas[a], Lorenzo Bottaccioli[a,b,*], Elisa Guelpa[c,b], Vittorio Verda[c,b], Edoardo Patti[a,b]

[a] *Dept. of Control and Computer Engineering, Politecnico di Torino, 10129 Torino, Italy*
[b] *Energy Center Lab, Politecnico di Torino, 10129 Torino, Italy*
[c] *Dept. of Energy, Politecnico di Torino, 10129 Torino, Italy*

Abstract

District Heating (DH) technology is considered to be a sustainable and quasi-renewable way of producing and distributing hot water along the city to heat buildings. However, the main obstacle to wider adoption of DH technology is represented by the thermal request peak in the morning hours of winter days, especially in Mediterranean countries. In this paper, this peak-shaving problem is tackled by combining three different approaches. A thermodynamic model is used to monitor the buildings' thermal response to energy profile modifications. An agent-based model is adopted in order to represent the end-users and their adaptability to variations of temperatures in buildings. Finally, a Reinforcement Learning algorithm is used to optimally mediate between two needs: on the one hand, a set of anticipations and delays is applied to the energy profiles in order to reduce the thermal request peak. On the other hand, the algorithm learns by trial and error the individual agents' sensitivity to thermal comfort, avoiding drastic modifications for the most sensitive users. The experiments carried out in the DH network in Torino (north-west of Italy) demonstrate that the proposed approach, compared with a literature solution chosen as a baseline, allows to achieve better results in terms of overall performances and speed of convergence.

Keywords: Agent-Based Model, Demand-Side Management, District Heating, Peak Shaving, Reinforcement Learning.

*lorenzo.bottaccioli@polito.it

1. Introduction

District heating (DH) system represents an opportunity to reduce fossil fuel adoption and consequently the carbon emissions for space heating and domestic hot water consumption [1, 2, 3]. The management of DH is often not straightforward because of its large dimension and the large mass flow rates at stakes. The adoption of novel ICT (Information and Communication Technologies) solutions for the automatic control can be particularly useful to reach higher efficiencies [4, 5]. This is particularly important looking for optimal operations and fault detection [6, 7] while moving towards integrated management of different energy infrastructure [8, 9]. Novel ICTs have also been proven useful for optimization of Energy Internet [10] or optimal scheduling in renewable energy systems [11].

One of the biggest problems in district heating concerns the presence of peak in the building thermal demand that makes the overall thermal request strongly variable in time [12, 13]. This is particularly relevant in case the heat is supplied during the daytime and interrupted at night; this is done mainly because of the habits of the customers. In these cases, the thermal demand is particularly high in the morning in order to increase the temperature of the buildings and the heating circuits. This creates a morning peak in the overall thermal request of the DH system. The occurrence of peaks is undesirable mainly because they are covered with less efficient technologies and because they cause high mass flow rates and therefore bottlenecks in water distribution [14]. In the rest of this manuscript with the term *"power peak"*, we refer to the overall daily power peak, which often occurs in the early morning hours.

An interesting option to smooth the peak consists in modifying the schedules of the customer heating systems. This approach is mainly called Demand Side Management (DSM) or Demand Response (DR), since actions are taken on the demand, rather than the production, side. The name was born in the electric engineering field [15]. This operation should be done taking into account

carefully the various systems and stakeholders involved: i) operators require a profile as flat as possible; ii) customers require that the modifications performed do not affect the indoor conditions of their buildings; iii) the presence of a district heating network make the aggregation of the building demands much different than the simple total request of building, because of its complex dynamics due to various phenomena (e.g. thermal losses, mixing, front delay and thermal transients). In order to properly select the DSM strategy, the building thermal demand should be known. In this context, a proper model for the selection of the best heating schedules of the customers connected to a district heating can be very useful; however in order to be effective, this should include the three above-mentioned aspects.

1.1. Related Works

In the literature, various attempts have been undertaken in this direction. A simulation on an established building has been performed to find the optimal DSM strategy from a building owner's perspective in [16]. In the case analysed power peak shaving larger than 30% can be achieved without significantly affecting indoor conditions. An agent-based intelligence has been used in [17] to flatten the thermal load of a group of buildings. A Stable Roommates algorithm has been adopted in [18] in order to minimise the sum of the thermal request of 28 buildings in England. In all these cases the effects of the network dynamic in the formation of the overall DH thermal request was not taken into account. In [19], an optimisation through genetic algorithm has been performed with the aim of minimising the maximum peak value. Results show that a reduction of 10% can be achieved in the overall thermal demand with very small schedule variations. In [20] the same algorithm was used to find the best rescheduling, including more significant modifications, while taking into account the effects of the indoor temperature changes. In [21] a field test campaign on two District Heating networks has been presented, adopting the STORM controller, which enabled a peak reduction in the range of 7.5–34%, saving operational costs and

CO_2 emissions. In [22] an active control strategy, based on a Model Predictive Control algorithm, is developed to maximise the profit of a cogeneration plant, by using buildings as storage capacity and selling electricity on the spot market when the price is the highest.

An optimisation technique such as Reinforcement Learning (RL), instead, seems to have been employed only marginally for directly tackling the peak demand problem in district heating, as its applications to peak-shaving mostly belong to the context of electric energy. In [23] an iterative Q-learning algorithm is used for energy arbitrage and peak-shaving of thermostatically controlled loads connected to a district heating system. In [24], the authors use temporal-difference learning to optimally control residential energy storage systems for minimising energy cost.

On the other hand, the optimisation of district heating systems has been approached several times with the help of agent-based technology. In [25], substations at the buildings level are equipped with reactive agents, which monitor consumes and make predictions about future energy requests, with the objective of general efficiency improvement of the system. Thermal load management at the buildings level is tackled in [26] with the purpose of reducing the thermal peak at the individual building level, while maintaining the thermal comfort for the users. Demand-side management through multi-agent based models, which coordinate individual requests in order to reduce the intra-daily fluctuation in thermal request, is also adopted in [27, 28, 29].

1.2. Proposed Contribution

Our work aims at optimising customers' heating systems startup times, rescheduling them, with the aim of minimising the thermal peak load without affecting end-users comfort. This is carried out with a tailored approach that allows to find the optimum taking into account both the customer satisfaction level and the district heating network dynamics. In this perspective, a satisfactory answer to the peak shaving problem would have to shave significantly the morning peak thermal request while respecting the thermal comfort of end-

users and providing a realistic thermodynamic model to simulate a real-word district heating network. With respect to previously cited literature works, our approach aims at providing a novel, more realistic and organic solution that encompasses all above-mentioned aspects.

- The thermal request peak is flattened by a Reinforcement Learning algorithm, which regulates the behaviour at the central plant level and is responsible for rescheduling the startup times in the heating network.
- An agent-based model provides the learning algorithm with feedback about the response to thermal changes of individual end-users, in order to guarantee that customers thermal comfort thresholds will not be exceeded as a consequence of the rescheduling of the heating network startup times.
- The thermodynamics of the network is simulated using an already established technical tool, developed in [30].

The combination of these three methods represents a novel and more effective approach in the peak-shaving problem for DH systems. Furthermore, the solution presented in this paper unifies both the demand and the operator sides of the problem. The Reinforcement Learning agent is responsible for managing strategically the distribution of hot water at the central plant level, while the agent-based model acts on the demand-side level of the system, giving real-time feedback about the effectiveness of the strategies adopted.

The rest of this paper is organised as follows. Section 2 presents the proposed methodology to address the peak-shaving problem in district heating systems. Section 3 discusses the experimental results. Finally, Section 4 provides our concluding remarks.

2. Methodology

The problem will be tackled by combining three different models, as shown in Fig. 1: the *network model*, the *end-users model* and the agent at the *central plant level*. The network model is based on a technical tool presented in [30],

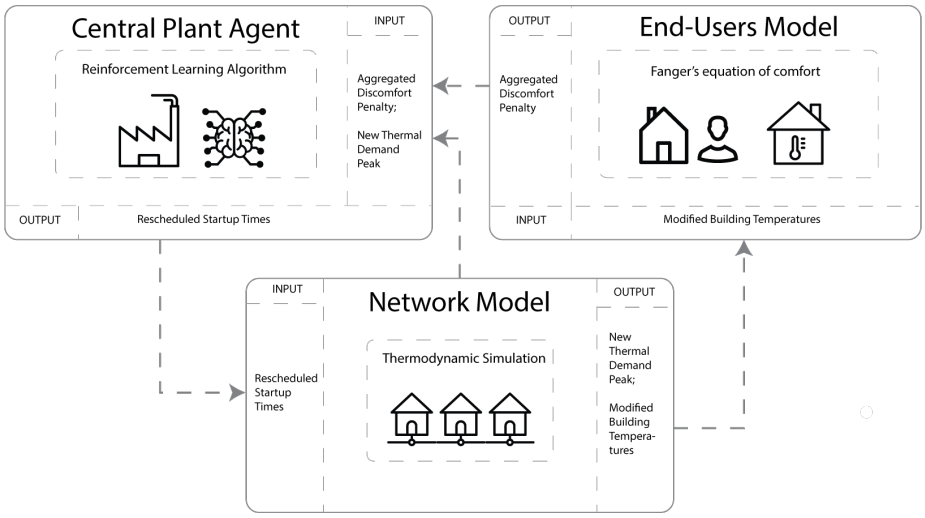

Figure 1: A schematic of the overall functioning of the presented system

it performs two tasks: i) computing the new aggregated energy profile at the cogeneration plant level given a set of startup times for the buildings in the network and ii) modelling the individual thermal responses of the buildings caused by the heating startup times rescheduling. These thermal modifications are then received as input by an agent-based model, which monitors the end-users' thermal comfort modifications caused by said variations in the indoor temperature of the buildings. Finally, a central agent based on a reinforcement learning algorithm is responsible for distributing the heating water throughout the network, mediating in the trade-off between the peak-shaving task at the cogeneration plant level and the individual comfort demands at the users' level, which jointly compound the reward, the main input of the reinforcement learning algorithm. Such a comprehensive threefold approach has been chosen because the effects of demand response, intended as anticipations and delays of the thermal request of buildings, affects several areas and specifically: i) the buildings, because of the change of the indoor conditions ii) the network, because of the different mass flows and temperature waves propagating in the pipelines iii) the central plant, which receives a different thermal load evolution. The adopted strategy is a day-ahead optimisation that exploits forecast thermal loads. This optimisation

aims at minimising the peak demand value; the independent variables are the schedule variations of heating systems (i.e. the startup time) for each building while the constraints are the maximum anticipations and delays. After the central plant agent completes a learning update, it provides the network model with a new set of anticipations and delays to be applied to the buildings, and the cycle starts over.

The rest of this section provides an in-depth description of each component shown in Fig. 1.

2.1. Network Model

The algorithm includes a model for the simulation of the DH network. This is done with the aim of taking into account: i) thermal losses during the water path; ii) the mix of various mass flow rates exiting the customer heat exchangers; iii) the thermal transients. In fact, if the pressure waves propagate within the entire network in few seconds, the temperature front propagates with a velocity comparable to that of the fluid flow. Therefore when the heating systems are switched on, the thermal plants (that receive large amount of mass flow at low temperature) supply a peak load that strongly depends on the thermal transient occurring in the network. The model used relies on a representation of the network as a series of pipes interconnected among them trough junctions. This is represented topologically as a set of branches interconnected among them through nodes. The topological representation is used for interconnected equations applied to the various branches. A suitable way to describe the network topology consists of using a matrix [31]. In our model, we defined a matrix A, where each value of the matrix a_{ij} is *0* if node i and branch j are not interconnected; *1* if i is the inlet node of the branch j; *-1* if i is the outlet node of the branch j. This matrix is built starting from information about the inlet node, the outlet node and their coordinates for each pipe. The adopted model includes the equations to compute both the mass and the energy conservation, which are applied to all the pipelines in the system. The former estimates the velocity and consequently the mass flow rates since density and sections are

available. The latter estimates the temperature field. The mass conservation equation is applied to all the branches in the network and the energy equation to all the nodes, considering thermal losses and neglecting thermal diffusion. The application to the entire network is done by relying on the incidence matrix as presented in [32]. The problem is solved to evaluate both matrices G and T, respectively including the mass flow in each branch and the temperatures in each node. This is done by taking into account proper boundary conditions, thermal losses, thermal transient and mass flow mixing/splitting.

The network model takes as input a data-set indicating the energy profile of a particular day for a series of buildings, and outputs the aggregate energy consumption at the level of the cogeneration plant (see Fig. 2). It is worth noting that the impact of possible estimations or forecasts of building thermal profiles on DSM strategies is moderate when the thermal load prediction is performed using data available from the thermal substation [33]. The initial energy profile data-set is a $T \times N$ matrix, where T is the number of time-steps in which the individual energy profile is divided, and N is the number of buildings in the data-set. Furthermore, the model also enables the monitoring of the indoor temperature shifts in buildings, caused by modifications in the energy profiles produced by the central agent's actions. Detailed information on the network model is provided in [32].

The model allows the agent at the central plant level to perform actions on the individual building, modifying its respective energy profile. The set of possible actions is discrete and comprises four anticipations (10, 20, 30 and 40 minutes), two delays (10 and 20 minutes) and a neutral action. The energy profile between 4 AM and 9 AM of each building can be anticipated or delayed accordingly as shown as example in Fig. 3, producing an alteration in the indoor temperature of the building, which is calculated by the thermodynamic model and used as explained below to evaluate the thermal comfort of users, who occupy the buildings.

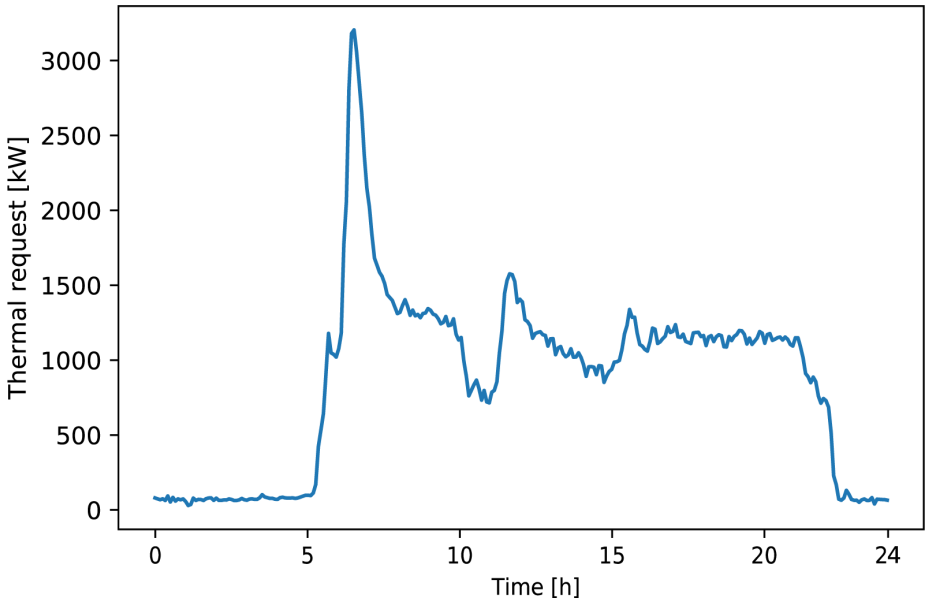

Figure 2: Aggregated energy profile as generated by the network model

2.1.1. Building Model

A compact building model has been used to evaluate the modification of the indoor temperature caused by the modification in the heating schedule. The model includes the thermal substation and the entire building, simplified as a unique equivalent room. This means that an average indoor temperature T is considered for the entire building. The building is modelled with an energy balance (Eq. 1) that includes the unsteady term, the heat provided by the heating system and the losses with the environment.

$$Mc_p \frac{dT}{dt} = \Phi_{syst} + \Phi_{loss} \tag{1}$$

where the left-hand side term is the unsteady term (related to the indoor temperature variation), Φ_{syst} is the thermal power provided by the heating system and Φ_{loss} the thermal losses towards the environment. The losses are expressed through a global dispersion coefficient per unit volume. This is evaluated considering the afternoon steady state when the power provided is almost constant and the heating power is equal to the thermal losses (Eq. 1 reduces to the right-

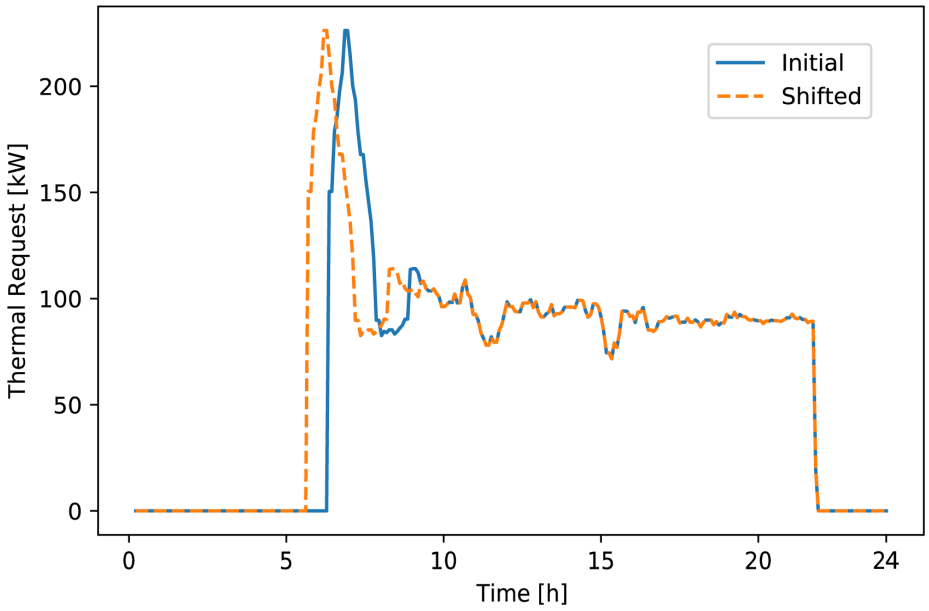

Figure 3: Example of energy profile shifting in a building (in blue the original profile, in orange the proposed shift). Note that the shift can only take place between time-steps 4AM and 9AM.

hand side terms). The indoor temperature can be estimated as approximately at 20°C while outdoor temperature and thermal power provided to the system are measured by the control system installed in the substations. The estimation is done for each day of the available data (for the same afternoon the outdoor conditions are supposed to not change significantly) and the average value is used as the global dispersion coefficient per unit volume. The thermal capacity of the building is estimated considering the thermal transient that occurs after the switching off time. During this transient the unsteady term is equal to the thermal losses. If no indoor temperature are available it is possible to assume that the night thermal transient of the return water in the secondary circuit in the last part follows the transient of the building air. This means that a measured temperature of the return water in the secondary circuit can be adopted for the calculation of a lower bound approximation for the thermal capacity that allows conservative estimation of the indoor temperature changes. The

substation is modelled by using an effectiveness-NTU method and a time delay is considered to account for the average time necessary for the water circulation in the building circuit.

2.2. The agent-based model

An agent-based approach is adopted to model the individual users' reactions to thermal changes in buildings. Agent-based modeling has been proven useful in many scenarios for simulating real-life dynamics. In general, agents are autonomous entities that interact with one another giving rise to complex high-level phenomena, which might or might not amount to the simple aggregation of the individual behaviors. In this work, agents are identified with end-users inhabiting the buildings. Their task is that of providing the cogeneration plant with feedback about their thermal comfort. For simplicity, it is assumed that only one agent is associated with each building. For the sake of clarity, in what follows the individual agents composing the network will be mostly referred to as "end-users".

2.2.1. End-users thermal comfort

In order to quantify the variation of comfort in the end-users due to a thermal modification, the widely-used Fanger's equation is adopted [34]. Fanger's comfort equation (Eq. 2) puts in correlation human endogenous variables and environmental variables, with the thermal comfort perceived by the individuals.

$$M - W = H + E_C + C_{Res} + E_{Res} \tag{2}$$

Where M is the metabolic rate, W is the external work, H is the dry heat loss from convection and radiation E_C is the skin evaporative convecting heat exchange, C_{Res} and E_{Res} are the respiratory convective and evaporative heat exchange. From this equation follows the one defining the PMV value:

$$PMV = (0.303 \cdot e^{-0.036 \cdot M} + 0.028) \cdot L \tag{3}$$

where L is any unbalance in the previous equation:

$$L = (M - W) - (H + E_C + C_{Res} + E_{Res}) \tag{4}$$

Thanks to the thermodynamic model and Fanger's equation, it is possible to monitor the variation of comfort experienced by the end-users as a direct consequence of the thermal modification in buildings caused by a slightly anticipated or delayed heat provision. The equation estimates two related values: the Predicted Mean Vote (PMV) and the Predicted Percentage of Dissatisfied (PPD). While the first is a scalar metric of evaluation for thermal comfort (i.e. too cold or too hot), the second measure predicts the percentage of the standard population that would be dissatisfied in such thermal conditions.

These two metrics are correlated as such: a neutral judgement, the most comfortable condition, corresponds to 0 PMV and 5% PPD. Parallel to the increase of PPD, judgements also vary, going from the neutral assessment to stronger evaluations such as "warm" or "hot". According to the latest ASHRAE Standard [35], the band for thermal comfort is set to be between -0.5 and +0.5 PMV, roughly corresponding to a maximum of 10% PPD. This is the value interval that has been considered for the evaluation of comfort in this experiment.

For the sake of simplicity, it can be assumed that people are generally satisfied with the thermal conditions of their buildings - otherwise, they would modify them until they reach a satisfactory situation. However, this is not always the case in real-life scenarios as there might be people that are more or less satisfied with their initial thermal situations. In this case, a modification of the indoor temperatures might worsen, but also better, their comfort. This is naturally due to the fact that a thermal variation can warm up the building of somebody who is experiencing discomfort due to cold and vice-versa.

In our study, however, any thermal variations will be considered detrimental to the end-user comfort. In general, a person is in thermal comfort in the range

between -0.5 and +0.5 PMV, which, according to the ASHRAE standard and Fanger's equation, happens when the thermal modification is not bigger than roughly $\pm$ 2.3°C. Any thermal modification bigger than $\pm$ 2.3°C will make the users closer to the PMV comfort limit of $\pm$ 0.5. By assuming that all agents start the simulation in an optimal thermal situation, then, we guaranteed that *any* thermal variations whatsoever in their buildings will be detrimental to their comfort levels.

Since everybody reacts differently to thermal modifications in their environments, every end-user is modelled as having different sensibilities towards thermal changes. This is achieved by having different threshold values for thermal comfort for each different agent. A Gaussian distribution with mean 2.3°C and standard deviation of 0.5°C models this set of thresholds. By doing so, it is possible to describe a population of end-users that will have different thermal sensibilities, as some will be more resilient to temperature variations, while others will exceed the PMV comfort range of $\pm$ 0.5 much more quickly.
Whenever end-users will experience discomfort due to thermal modifications bigger than their individual threshold, a discomfort penalty d is calculated and sent to the central agent:

$$d \doteq \sum_{n=1}^{N} P_n \tag{5}$$

where

$$P_n \doteq \begin{cases} max(|\Delta T_n|) - K_n, & \text{if } max(|\Delta T_n|) > K_n \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

ΔT is a matrix of temperatures describing the daily variations of temperature for each building n in the network, caused by the central agent's activity. The max is taken in order to ensure that the discomfort is calculated in the worst possible case, considering the largest difference per day for each user between the initial and the modified situations. K represents the thermal comfort thresholds (in °C) of each individual end-user. d is non-zero only when the variation of temperature, in absolute, is larger than the comfort threshold for at least one particular end-user.

End-users are therefore modelled as different from one another, that is, each and every individual has a different propensity to accept thermal changes, and will react accordingly. When a users' thermal variation exceeds their individual threshold, a real-time penalty is calculated and sent to the central agent at the cogeneration plant, which will then learn to avoid further penalties by understanding which users are more sensitive. A successful algorithm should then comprehend and exploit this heterogeneity in the population's adaptability to thermal changes, by, at first, discovering each user's thermal sensibility and, secondly, exploiting this piece of information by adopting more radical actions on the more resilient users, while being softer with the more sensitive ones.

In conclusion, an individual user observes the temperature modification of its building caused by the action of the learning agent at the cogeneration's plant level; then, it assesses if the resulting difference in temperature is above its own comfort thermal threshold; if it is, the user signals the discomfort, which is ultimately added up for the whole network of end-users.

2.3. The central plant agent

This section presents the central plant agent and its input. Figure 4 presents the working flow of the central plant agent and the interactions with the other models in the system.

The process can be described as follows: the initial energy profiles E, a $N \times s$ matrix, specifying the initial energy profile for a particular winter day for each of the N users, divided into s timesteps, is mapped by the Network Model into their aggregation, and the initial peak P is obtained. Then the Central Plant Agent provides a set of strategies A to be applied to the initial energy profiles E, in order to get a new set E' and a relative new peak P'. A is represented by a matrix of dimension $N \times Q$, where Q is the number of possible actions per building. A penalty signal is now calculated by the Agent-Based Model, which simulates the end-users reaction to the thermal changes occured in their building as a consequence of the rescheduling of the startup times produced by the set of actions A. The reward R is now calculated (Eq. 7) from the observed

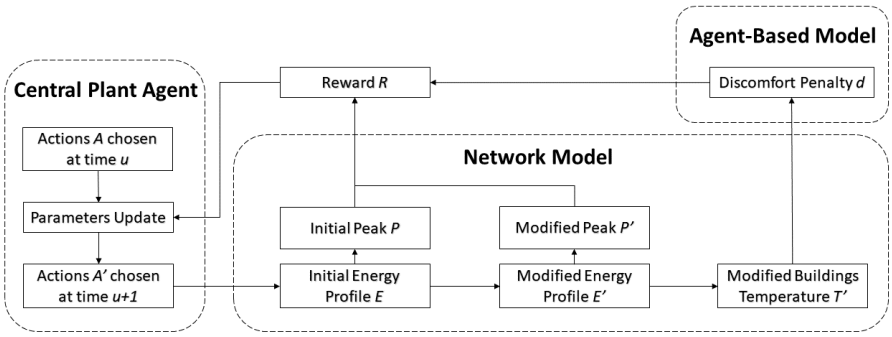

Figure 4: Work-flow diagram

reduction of the thermal request's peak and the possible penalty signals gathered from the network of users (Eq. 5). The cogeneration plant's agent then receives the reward and, considering the set of actions that produced that reward, it updates its learning parameters, selecting a new set of actions A' which is again applied to E. Through this process the Central Plant Agent learns how to effectively reduce the thermal peak while preserving the individual end-users' thermal comfort.

2.3.1. The reward

The focus of the present work, as said, is that of finding an optimal solution to the peak-shaving problem. The objective of the optimization problem is that of maximizing a reward as calculated by a reward function, which has to encompass all the requirements that a strategy should satisfy to be optimal. More specifically, the goal of our task is that of reducing the morning peak in the thermal request at the cogeneration plant level without discontenting the end-users. On the one hand, then, the reward has to increase if the morning peak is reduced; on the other hand, the reward has to reduce if thermal discomfort is induced into the end-users. To accurately model these two aspects, the reward

has been calculated as follows:

$$R \doteq (P_{init} - \mu_{plateau})/(P_{new} - \mu_{plateau}) - c \cdot d \tag{7}$$

where P_{init} and P_{new} are respectively the old and the new maxima (their peaks) of the old and the new modified aggregated thermal profile, in kW, normalised by subtracting $\mu_{plateau}$, the average value of the afternoon energetic plateau. d is the discomfort penalty as previously shown in Eq. 5. Constant c ($\frac{1}{^{\circ}C}$) is responsible for mediating between the peak shaving task and the discomfort of users: a lower value will favor the maximization of the peak reduction, conversely, a greater value gives greater importance to the preservation of users comfort levels. This constant can assume any real value but has been experimentally found to be helpful in balancing out the two sides of the reward function when equal to 1. In this setting, any trespassing of the users' comfort thresholds is greatly penalized as very detrimental to the overall reward.

2.3.2. Reinforcement Learning: Gradient Bandit

In this paper, it is argued that a better way to solve the peak-shaving problem is that of modeling it as a Multi-Armed Bandit (MAB) scenario, and solving it with the help of a reinforcement learning algorithm. A MAB is generally described as a game at a slot machine game (this is why a "bandit") with multiple levers, each representing one of k possible outcomes. The objective of a MAB agent is that of finding the lever which leads to the highest expected average reward. A MAB problem is considered to be a "stateless" and thus simplest Reinforcement Learning scenario as the reward only depends on the action taken, independently by anything else. The MAB algorithm that has been chosen for the peak-shaving scenario is the Gradient Bandit Algorithm (GBA) [36].

Differently to other MAB algorithms, gradient bandit estimates the desirability of an action a by learning a numerical $preference$ $H_u(a)$ associated with it. The larger the preference for a certain action, with respect to the other possible actions, the more often that particular action will be chosen. The probability

π of choosing action a at timestep u is determined according to the soft-max distribution:

$$\pi_u(a) \doteq \frac{e^{H_u(a)}}{\sum_{b=1}^{k} e^{H_u(b)}} \tag{8}$$

In which a is the chosen action, while b is one of the total k actions, which are added up to normalize the numerator into a probability value. Initializing all preferences with the same value guarantees that every action will be chosen with the same probability at the startup of the algorithm. The update of the preference for action A at time u is done as follows:

$$\begin{aligned} H_{u+1}(A_u) &\doteq H_u(A_u) + \alpha(R_u - \bar{R_u})(1 - \pi_u(A_u)), \qquad \text{and} \\ H_{u+1}(a) &\doteq H_u(a) - \alpha(R_u - \bar{R_u})\pi_u(a), \qquad \forall a \neq A_u \end{aligned} \tag{9}$$

where $\bar{R_u}$ is the current average of rewards obtained that far, which serves as a baseline for the immediate reward R_u, stabilising training and reducing the variance in the updates. The variable α instead represents the learning rate (or step size), which can assume any real value, preferably in the set (0,1]. Bigger values of α will make the agent converge faster towards a solution; however, very large steps make it less probable that the found solution effectively approximates the global optimum of the problem.

Among all MAB algorithms, gradient bandit has been mainly chosen because it allows the update of the utility values associated with each action, thus their probabilities, to be performed simultaneously for each building present in the network. In general, in fact, MAB algorithms are devised to tackle simpler problems in which one and only one action can be considered at once [36]. This would make them hardly applicable to our case as, in our use-case scenario, the agent at the cogeneration plant level is required to simultaneously update the actions' preferences for all buildings at the same time. The overall amount of possible sets of actions that can be taken in our case is in the order of N^Q different possibilities, which makes practically impossible to apply a different MAB algorithm. On the other hand, with the Gradient Bandit Algorithm,

once a reward is gathered from the thermodynamic model, the central agent can directly update the whole preference table, indicating the utility value for each of the N actions taken and each of the remaining $(N \times (Q-1))$ unchosen actions. Another reason to choose the GBA is that it has also been proven to be an instance of stochastic gradient ascent, a well-known optimisation method with very desirable convergence properties [36]. In a nutshell, such properties ensure that each step taken by the algorithm goes into the direction of the best possible solution, even though this might be reached at approximately infinite time, depending on the complexity of the task. The fact that gradient bandit relies on a stochastic method to approximate the best solution is a convenient characteristic, since the high dimensionality of the problem investigated implies the need for a very thorough exploration of the action-space.

Algorithm 1 Gradient Bandit: single day scenario

inputs: initial thermal profile E, initial aggregation P and initial indoor temperatures T, comfort threshold K and maximum time i_{max}
initialise preferences $H_{u=1}(A) \leftarrow 0$, cycle counter $u \leftarrow 1$ and initial time $i_{init} \leftarrow$ *current time*
$i \leftarrow 0$
while $i < i_{max}$ **do**
    $u \leftarrow u + 1$
    update $\pi_u(A_u)$ according to equation 8
    with probability $\pi_u(A_u)$, sample set of actions a_u
    apply a_u to E to get the new profile E'
    feed E' to model M and get new aggregated profile P', and temperature variation T'
    calculate penalty d according to equation 5
    obtain reward R according to equation 7
    update $H_{u+1}(A)$ according to equation 9
    update the step size α
    $i \leftarrow$ *current time* $- i_{init}$
end while

In our experiments a second scenario is also considered. In this scenario, every 2 hours of training, the initial energy profile is updated and it is set to be the best energy profile found by our central plant agent in such 2 hours interval.

Concurrently to the update of the initial energy profile, the preference table is reset to zero in order to facilitate the training on the new profile without biases, as shown in Algorithm 2. The main reasons for including this second scenario are more clearly detailed in Section 3.2.

Algorithm 2 Gradient Bandit: iterative scenario

inputs: initial thermal profile E, initial aggregation P and initial indoor temperatures T, comfort threshold K, time between reset i_{reset} and maximum time i_{max}
initialise preferences $H_{u=1}(A) \leftarrow 0, R_{best} \leftarrow 0, E_{best} \leftarrow E$, cycle counter $u \leftarrow 1$ and starting time $i_{init} \leftarrow current\ time$
$i \leftarrow 0$
while $i < i_{max}$ **do**
    $u \leftarrow u + 1$
    update $\pi_u(A_u)$ according to equation 8
    with probability $\pi_u(A_u)$, sample set of actions a_u
    apply a_u to E to get the new profile E'
    feed E' to model M and get new aggregated profile P', and temperature variation T'
    calculate penalty d according to equation 5
    obtain reward R according to equation 7
    update $H_{u+1}(A)$ according to equation 9
    update the step size α
    if $R > R_{best}$ **then** $E_{best} \leftarrow E'$
    end if
    $i \leftarrow current\ time - i_{init}$
    if $i >= i_{reset}$ **then** $u \leftarrow 1$ $E \leftarrow E_{best}$ and $H_u(A) \leftarrow 0$
    end if
end while

3. Experimental Results

Our solution, based on the Gradient Bandit Algorithm (GBA) above presented, is evaluated and compared with a Genetic Algorithm (GA), taken as a baseline for its proven satisfactory results in the peak-shaving problem context [30]. Genetic Algorithms are powerful methods of optimization [37], based on the subsequent generation of population of strategies, which are iteratively evaluated and selected according to a fitness score (the reward, in our case). The

two methods will be evaluated on two scenarios. In the first one, the basic case is considered, where the optimisation is always performed on the same initial energy profile. In the second case, the initial profile is updated, applying actions to it iteratively. The criteria upon which a strategy is evaluated are the overall peak reduction and the respect of the thermal comfort of users. A satisfactory result would simply be one in which a significant peak reduction (i.e. larger than 30%) comes along with the respect of most users' thermal thresholds. In general, a very good improvement in the overall thermal peak load, combined with a poor result regarding the comfort of users, would be deemed as a failure.

The initial energy profiles used in the experiments correspond to real data gathered on a winter day in the city of Torino, which hosts the largest district heating network of Italy. Such demand profiles have been provided by the local heating provider.

The thermal comfort thresholds K for the $N = 66$ end-users in the network are randomly sampled once, before the experiments start, and are kept constant across them so that the methods tested could be compared upon the same users' conditions and sensitivity. The overall results obtained at each trial strongly depend on the way in which the users' thermal thresholds are modeled, so it is important to keep them constant. The results can indeed be slightly better if the thresholds are more permissive, and, vice-versa, slightly poorer. However, the comparison of the proposed methods and the general dynamics of the outcome are not affected by a particular choice of thresholds if these are held constant across the experiments.

Since genetic and gradient bandit algorithms have a very different training process, with heterogeneous components, the two will be tested against each other using as a common scale their computational times. These are calculated on the same machine, without using parallelization or any other time-optimization technique during the test of the algorithms. Even though our solution and GA work differently, they are both based on one main training loop, which roughly takes the same amount of seconds to be computed. More specifically, the ra-

tio between individuals in a generation for GA and preferences update in our algorithm is roughly 1:1. It is fundamental to compare the methods against their relative times because one approach might yield the same results as the other one but in a significantly smaller amount of time, making it an overall better technique for a real-life application. In fact, the possibility of quickly retraining the algorithm, adapting it to several different startup conditions, is indeed a very desirable property, which would allow a wide-spread application of this method to a variety of real-life scenarios.

The server's CPU on which all the experiments have been carried out consists of 2 Intel(R) Xeon(R) E5-2630v3 @2.40GHz CPU (8 cores, 16 threads each). As said, the data-set taken as input consists of 66 energy profiles, divided into 288 5-minutes time-steps.

3.1. Single step scenario

This is the standard case, as presented in the previous section. Our proposed algorithm and the Genetic Algorithm train always on the same energy profile from the same winter day and, after each generation or preferences update, the input is reset to be the initial profile. A reward as calculated in Equation (7) is gathered at each step. Training time is 10 hours of computation for both methods, corresponding to roughly 200 generations for GA and 20,000 preferences update for our solution. In Fig. 5, the raw results for both methods are shown in terms of reward, and their hourly moving average is calculated. It can be seen that our solution maintains a better average and variance during the whole experiment. The higher instability for GA is due by the way it updates the strategies: many inferior strategies are able to persist over the generations despite being selected among the best 20%, while mutations are random and only rarely improve the fitness of the strategies. Our solution, instead, quickly approaches a local optimum, rapidly increasing the overall reward in the first handful of steps, and then converges around a few best solutions with smaller variance and in general much more consistency than GA. Since our solution is based on a Gradient Bandit algorithm, a stochastic method, some variance is

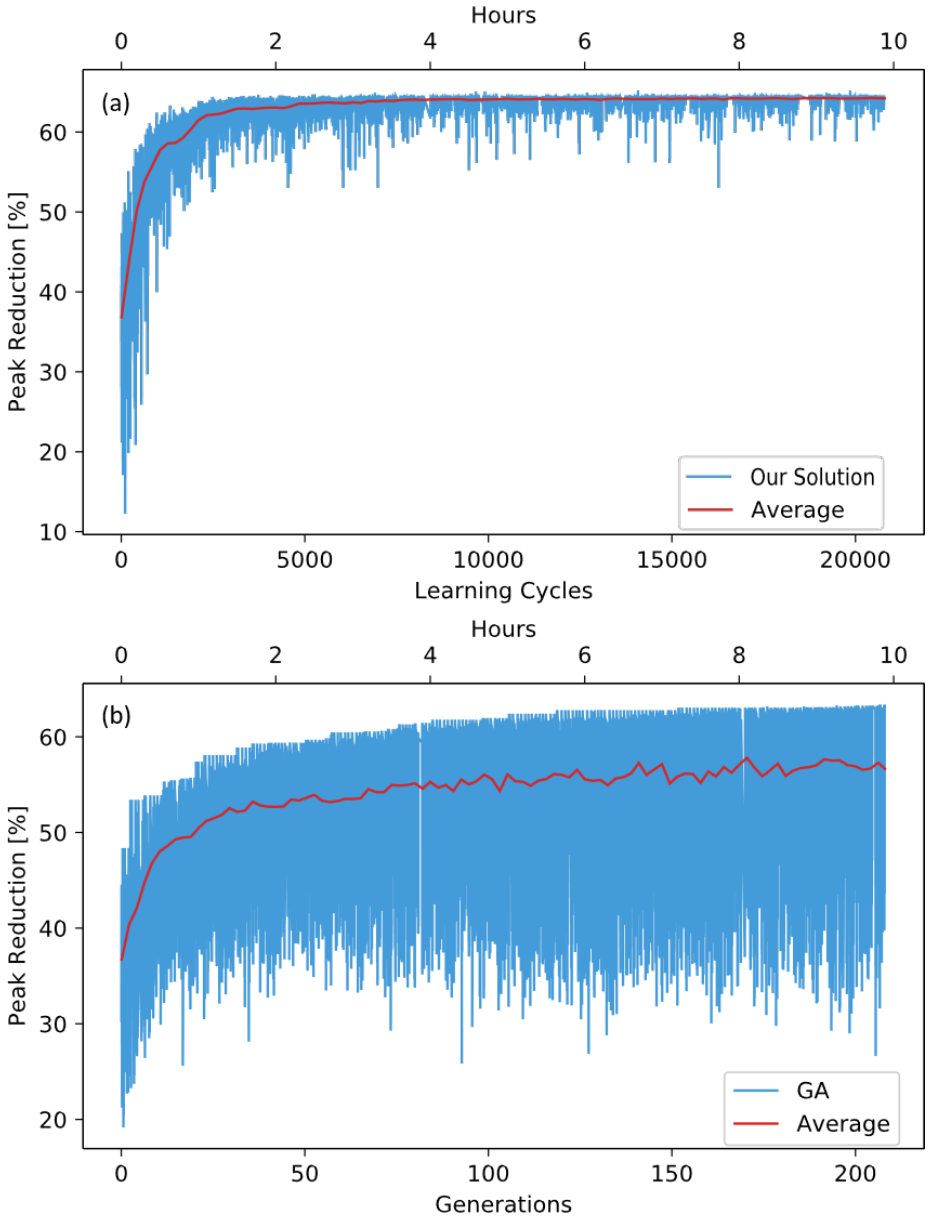

Figure 5: The overall results and the moving average per hour (a) for our solution, and (b) for GA.

inherent to the method itself and can only be reduced in time with a higher learning rate α, at the cost of a convergence to poorer results.

Average and variance show how well a strategy is performing throughout time, but are not indicative of the best results achieved by such a method. The purpose of this optimisation, however, is that of finding the strategy that better tackles the peak-shaving problem disregarding how many times such strategy appears during the simulation.

It can be argued that in order to be applicable to real-life scenarios, these methods need to quickly adapt and understand the environment at hand, finding the best way to quickly reduce the peak respecting the users' thermal comfort. In Fig. 6, the maximum reward obtained per hour is represented in the top graph. This is calculated by taking the best result obtained in each hour bucket of training time. Again, it can be observed that our solution consistently does better than GA at any moment of the simulation. To further illustrate this difference, the graph at the bottom compares the initial energy profile, in blue, with the profiles associated with the best strategies discovered by the two methods after the first 30 minutes of training. Our solution reaches a 63% reduction in the peak and GA remains under 55%. If in the long term their relative distance closens, in the first few minutes of training our algorithm shows much bigger improvements and a faster convergence rate. In order to achieve higher percentage of peak reduction, it is necessary to make slight and precise adjustments to avoid harming the thermal comfort of users while further reducing the peak. The reward, indeed, takes into account the penalty received by the network of users, therefore every attempt at improving the peak reduction cannot ignore the individual sensitivity of the users. GA seems to be much slower and limited at this task, as its high variance slower slope of the curve illustrate. While our solution is able to avoid heavy penalties after the first few iterations, GA, because of the way populations are formed, tends to incur in this problem even at later stages of this experiment.

The fast adaptability of the proposed solution can be greatly helpful in real-life applications, in case, for instance, that the energy provider wants to find fitted strategies on a daily basis, providing the algorithm with a new energy profile every day and expecting solutions to be discovered in a small amount of time.

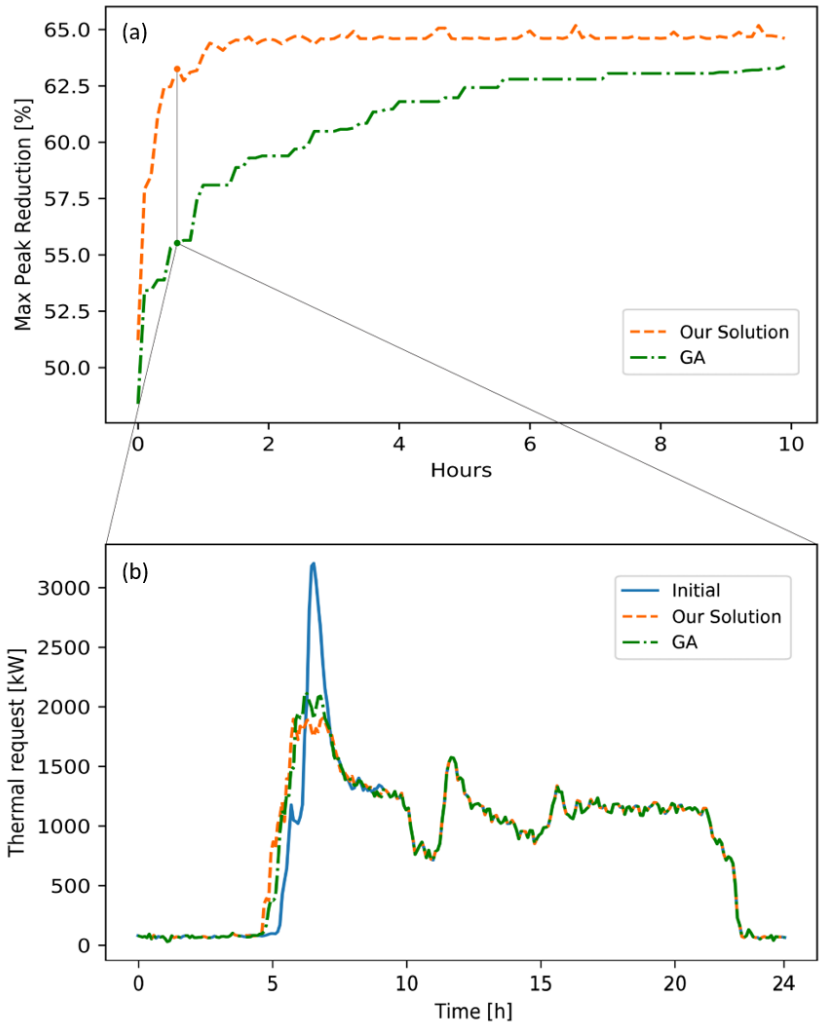

Figure 6: (a) The maximum reward per hour for our solution and GA, and (b) the peak-shaving dynamics for the best results after 30 minutes of training.

Summarizing, while both methods provide satisfactory results at the end of the 10 hours training on the simpler scenario, our solution converges to a higher result than GA, with much better performances, especially in the first couple of hours of training. Also, it is important to mention that the best strategies

discovered by both methods do not harm the thermal comfort of users, as the maximum temperature difference induced in buildings by these optimal strategies is well under their comfort thresholds.

3.2. Multiple steps scenario

In this scenario, everything works as before. However, every 2 hours of computation, the initial energy profile is substituted by the energy profile corresponding to the best reward obtained in that specific training interval. To be clear, consider the aggregated energy profile in orange of Fig. 6, produced by the strategy found by our solution. If after the first 2 hours that would be the best profile found, then from the 2^{nd} to the 4^{th} hour of training it would become the new initial profile to which actions are applied. After switching the profile, both algorithms reset, by sampling a new random population or by setting the preferences table to zero, causing the training to slow down in the first few iterations after a switch happens.This iterative scenario then outputs a series of sets of anticipations and delays, to be applied sequentially to the initial energy profiles.

One reason to include or even to prefer this approach is that it gives the learning agent at the central plant level much more freedom to apply anticipations and delays iteratively to the heating startup times of buildings, meaning that a stronger peak-reduction can theoretically be obtained. However, greater anticipations and delays can also mean that the end-users might suffer from higher thermal deviation and thus experience discomfort. This scenario gives then more space to improve over the previous results, at a possibly higher cost for the end-users thermal comfort. Fig. 7 shows the optimal strategies found after 10 hours of training for both methods. The difference, in this case, is much more evident than in the first scenario, because the iterative optimisation greatly enhances the gap between the convergence speeds of the two algorithms. On the one hand, our solution is able to level the peak almost perfectly, reducing it by around 80% of its previous height; on the other hand, there is still space for improvements for the strategy found by GA, which reaches a reduction of 75%.

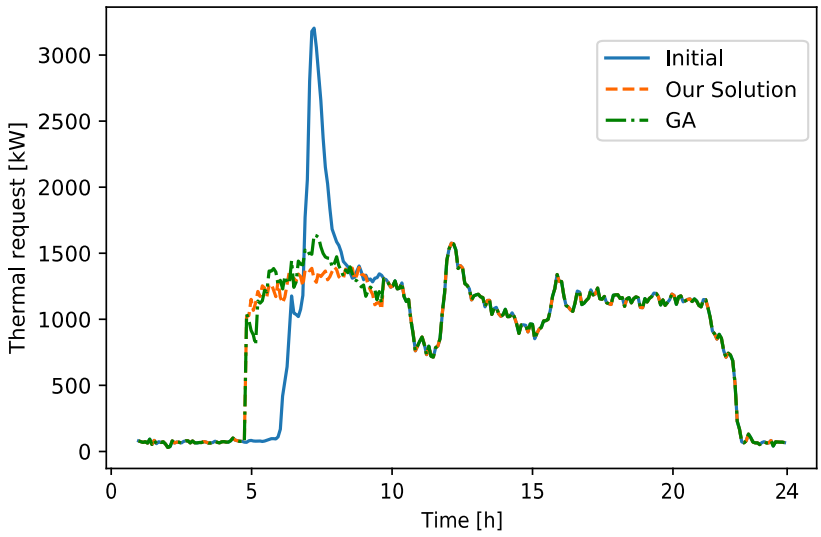

Figure 7: The best strategies found in 10 hours of multi-day training for our solution and GA.

This different ability for exploiting the multi-optimisation scenario can be observed more clearly in Fig. 8, where the maximum reward per 30 minutes is represented. In the first 2 hours, the difference is not significant, but it gets wider after successive updates of the energy profiles. This, as explained before, is due to the much faster convergence rate of the proposed solution with respect to GA, and its capability for quickly optimising under new conditions. It is worth nothing that, at each update, the chances of harming the users' thermal comfort increase, as it can be seen by the sudden drops in reward, which, as explained before, is a compound measure of the overall peak reduction and the comfort of end-users. Also note that, despite representing the best reward found every 30 minutes of training, the curve is not necessarily monotonic for GBA, because of the way its algorithm works. This happens because the stochastic gradient update performed in the GBA does not guarantee that, once reached, a so-far best solution is then always followed by a better one. In most scenarios, this is a desirable property that enables the algorithm to explore stochastically

the action-space, thus reducing the risk of "getting stuck" in a local optimum solution. The curve is monotonic however for GA because GA keeps a subset the best solutions across the development of successive generations, which guarantees that its best solutions are incrementing monotonically over time.

Our solution is more consistent in choosing strategies that respect the agents comfort thresholds, as it is able to train on the new profile much more quickly, thus having less drops in performances overall.

Fig. 9 shows a comparison between the best strategies found after 10 hours

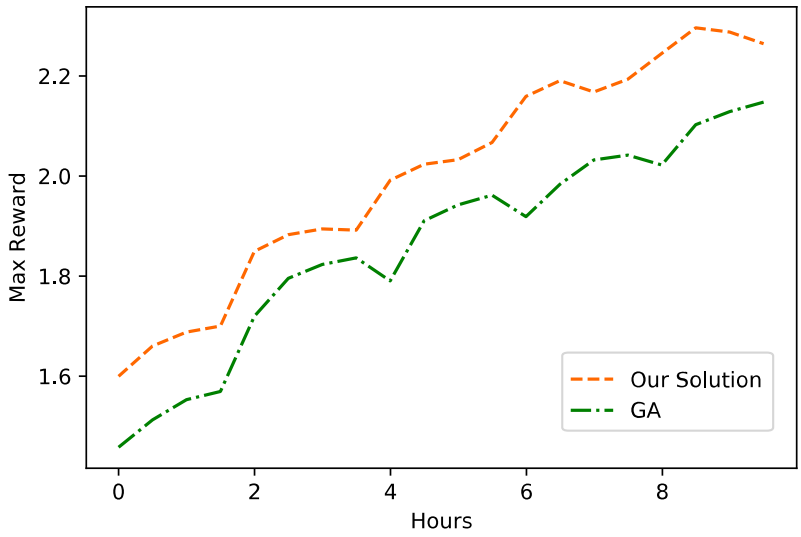

Figure 8: The maximum reward for our solution and GA in multi-step training, as calculated every 30 minutes.

of training for the single and the multiple step scenario. The greater freedom allowed in the multi-step case enables the Central Plant Agent to reduce the peak even more significantly.

Since in this scenario it is almost possible to reach the optimal solution to the peak-shaving problem, one might wonder if the thermal comfort of users is respected by the algorithm. Even very good results can be indeed invalidated if one or more users would be in thermal discomfort. Fig. 10 compares the

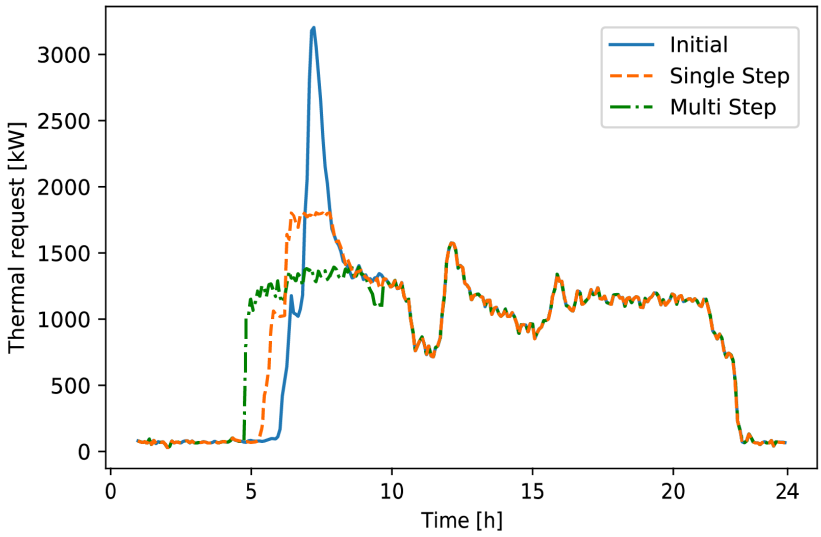

Figure 9: The best strategies found in 10 hours of training for the single step and the multi step scenario.

maximum temperature difference, per user, caused by the application of our solution in the second iterative scenario (shown in Fig. 7) and the thermal comfort thresholds of the users as modelled in these experiments. For each user, a lower value than the thresholds means that the user's thermal comfort has not been harmed, according to the chosen standard settings for thermal comfort [35], as explained in Section 2.2.1. It is remarkable how in the case of at least five users, the algorithm shows a significant capability for correctly identifying different kinds of users. For example, in three cases (i.e. building number 8, 25 and 52 in Fig. 10), the proposed algorithm exploits this knowledge by maximising the increase of the temperature in those buildings whose users are less sensitive. In two other cases (building 28 and 49 in Fig. 10), instead, an unusually low thermal threshold is met by a very small temperature increase. This shows a capability for learning the thermal sensitivity of each user, which is coupled with the thermal peak reduction at the aggregate level. It is even more remarkable considering that the thermal thresholds are arbitrarily sampled

from a random distribution and the algorithm has to adapt to them and to the thermodynamics of the model simultaneously. Anticipations, delays and thermal comfort are different components of a unique system and are strongly correlated, forcing the algorithm to learn how to combine all these factors and their consequences at the same time.

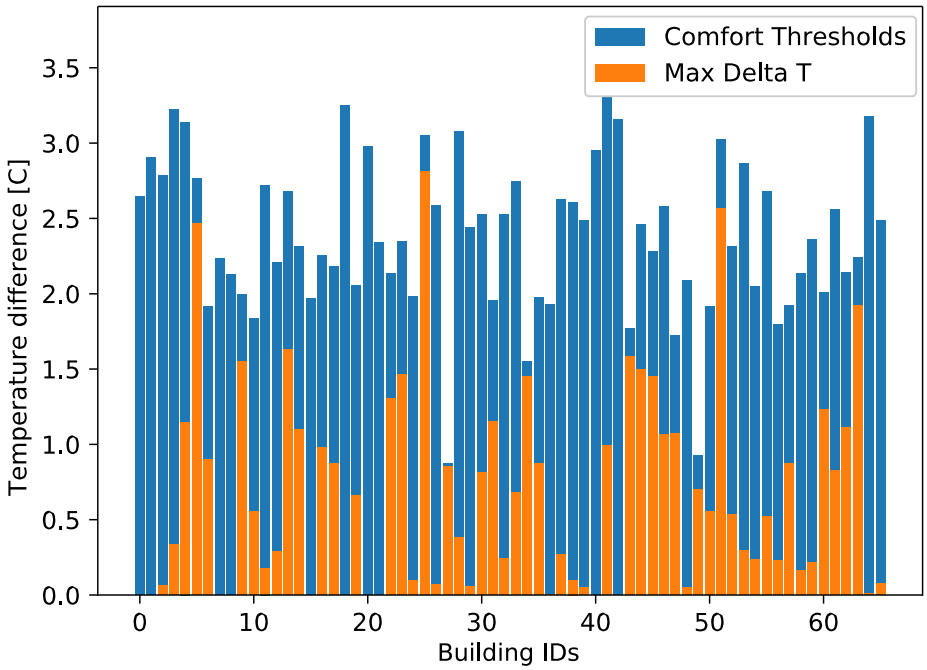

Figure 10: Maximum temperature difference per building (orange) caused by the best strategy for our solution and thermal comfort thresholds per building (blue).

In conclusion, the proposed technique allows to tackle effectively the problem described in this work. In particular, results have shown that, for the considered DH network, the proposed solution is able to outperform the Genetic Algorithm baseline in both overall results and speed of convergence. One reason for which the GBA is consistently faster than the Genetic Algorithm (GA) might be due to the way the two algorithms optimize the set of strategies. GA, after creating a population of different possible strategies, selects a set (around 10% in our case) of best performing strategies, which are randomly mutated and mixed

together, and added to new random strategies to form a new population to be evaluated. Doing so, GA also spends a lot of computational effort in creating and evaluating random and, thus, possibly inferior strategies. The Gradient Bandit Algorithm, on the other hand, associates a numerical value, the utility or preference, to each possible action that can be performed on every building, and then updates this value directly through the gradient bandit update rule. This enables for a great deal of computational saving compared to GA.

Even though both algorithms are able to mediate between the reduction of the thermal request's peak and the comfort of users, only our solution is able to fully exploit the properties of the users' network maximising the reduction of the peak under the constraint given by the users' thermal comfort. These results suggest that agent based modelling and reinforcement learning are mature technologies that can contribute to the optimisation of district heating systems in real-life applications, providing energy and monetary savings.

4. Concluding Remarks

In this paper, a novel approach to the peak-shaving problem in District Heating systems has been presented. Differently from the main method adopted in the literature [30], namely Genetic Algorithms, here the problem is tackled with the help of two different techniques. On the one hand, a Reinforcement Learning algorithm is responsible for modifying the energy profiles of buildings, mediating between the reduction of the thermal request peak at the central level and the temperature modifications at the buildings' level. On the other hand, a simple agent-based model monitors the internal thermal variations of the network of buildings, giving feedback to the central plant learning agent about the thermal discomfort that the end-users inhabiting the network might experience.

The purpose of this study was that of developing a method able to reduce the thermal request peak in morning hours, while respecting the thermal comfort of users. Carrying out two similar kinds of experiment, it has been found that

our proposed solution, based on a Gradient Bandit algorithm, delivers superior results in both scenarios compared to the Genetic Algorithm baseline, in terms of overall performance and speed of convergence.

There might be two, possibly coexisting, approaches to translate the methodology presented here into real-life applications. On one side, the whole process might be carried out offline, in a thermodynamic environment, which would have to be an accurate representation of the real District Heating network to be optimised. Furthermore, the agent-based model should represent faithfully the end-users of the real network, in terms of sensitivity to variations of temperature. This could be done by investigating the real end-users with questionnaires about their thermal sensitivity and by rewarding them if they agree to be part of such trial.

On the other side, primarily due to the difficulty in accurately modelling the end-users network, this process might take place online. In this case, the real end-users can provide the learning agent with live accurate feedback about their thermal comfort, allowing for a continuous online optimisation of the system. In order to avoid drastic falls in thermal comfort for the users during the early stages of the learning process, an intermediate setting is probably the best option: at first, the central plant agent is provided with an accurate thermodynamic simulation and a provisional agent-based model, then, after a few learning iterations, its best strategy is applied to the real environment and it is refined online, using live feedback gathered from the real end-users.

References

[1] A. Lake, B. Rezaie, S. Beyerlein, Review of district heating and cooling systems for a sustainable future, Renewable and Sustainable Energy Reviews 67 (2017) 417–425. doi:10.1016/j.rser.2016.09.061.

[2] M. A. Sayegh, J. Danielewicz, T. Nannou, M. Miniewicz, P. Jadwiszczak, K. Piekarska, H. Jouhara, Trends of european research and development in

district heating technologies., Renewable and Sustainable Energy Reviews 68, 1183-1192.

[3] H. Lund, B. Möller, B. V. Mathiesen, A. Dyrelund, The role of district heating in future renewable energy systems, Energy 35 (3) (2010) 1381–1390.

[4] M. Woźniak, k. Ksiażek, J. Marciniec, D. Połap, Heat production optimization using bio-inspired algorithms, Engineering Applications of Artificial Intelligence 76, 185-201.

[5] S. Farouq, S. Byttner, M. Bougueliaa, N. Nord, H. Gadd, Large-scale monitoring of operationally diverse district heating substations: A reference-group based approach, Engineering Applications of Artificial Intelligence 103492.

[6] S. Jakubek, T. Strasser, Artificial neural networks for fault detection in large-scale data acquisition systems, Engineering Applications of Artificial Intelligence 17(3),233-248.

[7] S. Delrot, T. Guerra, M. Dambrine, F. Delmotte, Fouling detection in a heat exchanger by observer of takagi–sugeno type for systems with unknown polynomial inputs, Engineering Applications of Artificial Intelligence 25(8),1558-1566.

[8] P. Mancarella, Mes (multi-energy systems): An overview of concepts and evaluation models, Energy 65, 1-17.

[9] E. Guelpa, A. Bischi, V. Verda, M. Chertkov, H. Lund, Towards future infrastructures for sustainable multi-energy systems: A review, Energy 184 (2019) 2–21.

[10] N. Zhang, Q. Sun, J. Wang, L. Yang, Distributed adaptive dual control via consensus algorithm in the energy internet, IEEE Transactions on Industrial Informatics.

[11] N. Zhang, Q. Sun, L. Yang, A two-stage multi-objective optimal scheduling in the integrated energy system with we-energy modeling, Energy 215 (2021) 119121.

[12] H. Wang, R. Lahdelma, X. Wang, W. Jiao, C. Zhu, P. Zou, Analysis of the location for peak heating in chp based combined district heating systems., Applied Thermal Engineering 87, 402-411.

[13] J. Ahn, S. Cho, Development of an intelligent building controller to mitigate indoor thermal dissatisfaction and peak energy demands in a district heating system., Building and Environment 124, 57-68.

[14] L. Brange, P. Lauenburg, K. Sernhed, M. Thern, Bottlenecks in district heating networks and how to eliminate them–a simulation and cost study, Energy 137, 607-616.

[15] P. Siano, Demand response and smart grids—a survey, Renewable and sustainable energy reviews 30, 461-478.

[16] K. Martin, Demand response of heating and ventilation within educational buildings, Master's Thesis, Aalto University, Espoo, Finland, 2017. Available online: https://aaltodoc.aalto.fi/handle/123456789/29149.

[17] H. Li, S. J. Wang, Load management in district heating operation, Energy Procedia 75, 1202-1207.

[18] T. Sweetnam, C. Spataru, M. Barrett, E. Carter, Domestic demand-side response on district heating networks, Building Research and Information 330-343.

[19] E. Guelpa, G. Barbero, A. Sciacovelli, V. Verda, Peak-shaving in district heating systems through optimal management of the thermal request of buildings, Energy 137 (2017) 706–714.

[20] F. G. Brundu, E. Patti, A. Osello, M. Del Giudice, N. Rapetti, A. Krylovskiy, M. Jahn, V. Verda, E. Guelpa, L. Rietto, A. Acquaviva,

Iot software infrastructure for energy management and simulation in smart cities., IEEE Transactions on Industrial Informatics 13(2), 832-840.

[21] T. Van Oevelen, D. Vanhoudt, C. Johansson, E. Smulders, Testing and performance evaluation of the storm controller in two demonstration sites, Energy 197 (2020) 117177.

[22] D. Vanhoudt, B. Claessens, R. Salenbien, J. Desmedt, An active control strategy for district heating networks and the effect of different thermal energy storage configurations, Energy and Buildings 158 (2018) 1317–1327.

[23] B. J. Claessens, D. Vanhoudt, J. Desmedt, F. Ruelens, Model-free control of thermostatically controlled loads connected to a district heating network, Energy and Buildings 159 (2018) 1–10.

[24] A. Ali, H. Kazmi, Minimizing grid interaction of solar generation and dhw loads in nzebs using model-free reinforcement learning, in: International Workshop on Data Analytics for Renewable Energy Integration, Springer, 2017, pp. 47–58.

[25] F. Wernstedt, P. Davidsson, An agent-based approach to monitoring and control of district heating systems, in: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, 2002, pp. 801–811.

[26] H. Li, S. J. Wang, Load management in district heating operation, Energy Procedia 75 (2015) 1202–1207.

[27] F. Wernstedt, P. Davidsson, C. Johansson, Demand side management in district heating systems, in: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, ACM, 2007, p. 272.

[28] C. Johansson, F. Wernstedt, P. Davidsson, Deployment of agent based load control in district heating systems, in: First International Workshop on AgentTechnologies for Energy Systems, Canada, 2010.

[29] F. Wernstedt, C. Johansson, Intelligent distributed load control, in: Proceedings of the 11th international symposium on district heating and cooling. Reykjavik, Iceland, 2008.

[30] E. Guelpa, V. Verda, Optimization of the thermal load profile in district heating networks through "virtual storage" at building level, Energy procedia 101 (2016) 798–805.

[31] P. Valdimarsson, Modelling of geothermal district heating systems, BHáskóli Islands, University of Iceland, Faculty of Engineering, 1993.

[32] E. Guelpa, A. Sciacovelli, , V. Verda, Thermo-fluid dynamic model of large district heating networks for the analysis of primary energy savings, Energy 184 (2019) 34–44.

[33] E. Guelpa, L. Marincioni, , V. Verda, Towards 4th generation district heating: Prediction of building thermal load for optimal management, Energy 171 (2019) 510–522.

[34] P. O. Fanger, et al., Thermal comfort. Analysis and applications in environmental engineering, Copenhagen: Danish Technical Press., 1970.

[35] R. American Society of Heating, G. Air Conditioning Engineers (Atlanta, ANSI/ASHRAE Standard 55-2017: Thermal Environmental Conditions for Human Occupancy, ASHRAE standard, ASHRAE, 2017.

[36] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[37] D. E. Goldberg, Genetic algorithms, Pearson Education India, 2006.