Coordinated Load Control of Renewable Powered Small Base Stations Through Layered Learning

Marco Miozzo, Nicola Piovesan, and Paolo Dini

Abstract—The massive deployment of Small Base Stations (SBSs) represents one of the most promising solutions adopted by 5G cellular networks to meet the foreseen huge traffic demand. The usage of renewable energies for powering the SBSs attracted particular attention for reducing the energy footprint and, thus, mitigating the environmental impact of mobile networks and enabling cost saving for the operators. The complexity of the system and the variability of the harvesting process suggest the adoption of learning methods. Here, we investigate techniques based on the Layered Learning paradigm to control dense networks of SBSs powered solely by solar energy. In the first layer, SBSs locally select switch ON/OFF policies according to their energy income and traffic demand based on a Heuristically Accelerated Reinforcement Learning method. The second layer relies on an Artificial Neural Network that estimates the network load conditions to implement a centralized controller enforcing local agent decisions. Simulation results prove that the control of the proposed framework mimics the behavior of the upper bound obtained offline with Dynamic Programming. Moreover, the proposed layered framework outperforms both a greedy and a distributed Reinforcement Learning solution in terms of throughput and energy efficiency under different traffic conditions.

Index Terms—Mobile networks, SON, sustainability, renewable energy, energy efficiency, machine learning.

I. INTRODUCTION

A T THE dawn of the modern telecommunication systems, when we entered the era of the smartphones, the main concern on the energy was related to the battery lifetime of the mobile devices. Nowadays, the power consumption has become an issue also for the Mobile Network Operators (MNOs). In fact, the high number of connected devices and the huge capacity demand of mobile services has provoked a large densification in the radio access infrastructure. Consequently, the energy cost has become an important portion of the Operational Expenditures (OPEX) of MNOs, e.g., it already reached the expenses paid for the personnel required to manage the network for a Western European MNO in 2007 [1]. Moreover, the forecast for the mobile traffic is of an annual growth rate of 53%, which will produce a further increase in the energy consumed by telecommunication infrastructures and, thus, on its footprint on the environment. The Information and Communication Technology (ICT) ecosystem was consuming 1500 TWh of energy in 2013 [2], which corresponds to the 10% of the worldwide electricity generation. Considering the mobile traffic trend, it is expected to reach up to the 51% in 2030 [3]. As a result, the energy sustainability is one of the key pillars in the design of the next generation cellular networks for OPEX and carbon footprint reduction. Many standardization bodies have already started working on this aspect, e.g., ETSI [4] and 3GPP [5].

In the last decade, the research community has been paying close attention to the Energy Efficiency (EE) of the radio communication networks. The dynamic switching ON/OFF of the Base Stations (BSs) has been identified as one of the most promising EE techniques, as testified by the rich literature [6]. However, this solution has been received with reserves from MNOs, since it might generate problems of coverage holes and possible failures of network equipment due to the frequent ON/OFF switches. Consequently, the introduction of Energy Harvesting (EH) capabilities represents an interesting approach to further increase the energy savings, allowing simultaneously to mitigate both the costs and the environmental impact of the mobile communication systems. Many studies proposed solutions using BSs powered with renewable energies [7], not only for regions with limited access to the power grid, e.g., rural areas, but also for deployments in urban environment. Furthermore, 5G will bring densified Heterogeneous Networks (HetNet) comprising a high number of Small BSs (SBSs) within the coverage of Macro BSs (MBSs), especially for satisfying the high traffic demand in urban scenarios [8]. The reduced energy requirements of the SBSs encourage the adoption of renewable energy for their power supply.

The reference renewable system for a SBS is composed of a Photo-Voltaic (PV) panel and a battery to store the surplus energy that cannot be directly used and make it available for the periods when PV source is not generating energy. The intermittent and erratic nature of the renewable energies has to be considered to manage the high variations in the incoming energy. In fact, even during the summer and in good weather conditions, the harvested energy in the peak irradiation hour can vary up to the 85%, as showed in [9]. Similarly, considering that the solar radiation intensity and the daylight duration vary significantly across the months [9], seasons have a strong

Manuscript received July 31, 2018; revised April 1, 2019 and July 16, 2019; accepted August 23, 2019. Date of publication September 2, 2019; date of current version March 18, 2020. This work was supported in part by the European Union’s Horizon 2020 Research and Innovation Programme through Marie Skłodowska-Curie (SCAVENTE) under Agreement 675891, in part by the Spanish Government through MINECO (5G-REFINE) under Grant TEC2017-88373-R, and in part by the Generalitat de Catalunya under Grant 2017 SGR 1195. The associate editor coordinating the review of this article and approving it for publication was H. Shan. (Corresponding author: Marco Miozzo.)

The authors are with the MONET Department, CTTC/CERCA, 08860 Barcelona, Spain (e-mail: mmiozzo@cttc.es; npiovesan@cttc.es; pdini@cttc.es).

Digital Object Identifier 10.1109/TCGN.2019.2938860
Learning solution and with realistic environmental conditions for both traffic load and energy arrivals, as will be also discussed in Section II. In particular, this work is the first to demonstrate that LL is a valid framework to efficiently manage mobile networks. Moreover, this is the first proposal, which uses HAMRL as the tool enabling the interactions between the algorithms at the different layers to improve agents’ coordination. As a result, the innovative contributions of this paper can be summarized as follows:

1) Design of an online optimization framework based on LL for the traffic control of densified HetNets powered with renewable energies tailored to an SDN/NFV architecture.
2) Characterization of the temporal behavior of the learned control policies.
3) Characterization of the performance of the LL framework and of the specific algorithms at each layer.
4) Comparison with an upper bound evaluated offline and with a distributed MRL scheme.

We note here that the LL solution has been previously introduced in our work [17]. However, the analysis has been carried out only in a limited scenario and without analyzing the temporal behavior of the learned policies and without the comparison against the offline solution.

The rest of this paper is organized as follows. In Section II, we present the state of the art in EE, with a focus on the online solutions. Section III defines the problem statement and the architecture of the control framework. Sections IV and V describe the solutions adopted in the two layers with details. Section VI is devoted to the presentation of the simulation scenario where the proposed approach has been evaluated, and to discuss the correspondent numerical results. Finally, Section VII concludes by summarizing the main achievements of the work.

II. RELATED WORK

In general, renewable energy systems are dimensioned to guarantee that the BS can autonomously and perpetually operate. However, in this case, the resulting PV panel sizes are large and not suitable for the deployment of SBSs in urban areas [18]. In [19], the authors relaxed this design assumption by introducing the concept of outage, defined as the fraction of time during which a BS cannot satisfy the traffic demand, due to energy shortage. The authors evaluate the size of harvesters and batteries as a function of the outage probability for different geographical locations. The main outcome is that the system may be feasible and cost-effective in locations with relatively high solar irradiation, considering the cost and dimension of the energy harvesting hardware and of the grid energy, when adopting an outage of 1% as design constraint.

Several offline solutions have been proposed for studying the behavior of EH in mobile networks, providing useful guidelines for the design of the network and their harvesting system. In [20], the authors presented a work on energy saving in $k$-tier HetNet with EH capabilities and sleep mode strategies. In this model, the authors used stochastic geometry to define a metric called availability $p_k$, which represents a fraction of time the $k^{th}$ tier of BSs has enough energy reserve to stay switched ON. The method provides the optimal values...
of the battery and the PV power capacity for the deployment of self-powered HetNets. Nonetheless, it considers a Poisson energy arrival process and continuous transmissions, which may provide a non-realistic approximation to these phenomena, as described in [9] and [21], respectively.

In [22], the authors introduced the concept of Zero grid Energy Networking (ZEN) for a mesh network of BSs powered only with RES. The rural scenario has been considered, where there is no connection to the electric grid and, therefore, the BSs need to be energy self-sufficient. The renewable energy system is designed by considering the daily typical traffic and energy harvested profile for the cities of Aswan, Palermo and Torino generated with a simulator called PVWatts [23]. In addition, the assumption of energy self-sustainability is relaxed and an algorithm enabling SBS sleep mode is introduced to reduce the dimension of the RES equipment. The proposed solution assumes that half the BSs in a given area can enter sleep mode when the traffic is below 50% of peak. The energy consumption and the PV panel dimension can be reduced by the 50%, but the effects on the service provided by the network are not evaluated.

The ski rental problem has been proposed in [24] to optimize the switch ON/OFF scheduling of EH SBSs. Using this approach, each SBS can decide to switch ON/OFF autonomously, without any prior information on future energy arrivals. The decision is performed every 10 sec., which implies that UEs might have to perform frequent handovers and the SBSs may be subject to numerous ON/OFF switches. The failure probability is high and cannot be ignored, since it represents the main constraint for MNOs to implement switch ON/OFF techniques in operative networks [25]. Moreover, authors in [26] show that only larger time quantization (e.g., 1 hour) can achieve a correct dimensioning of the solar power system for cellular base stations.

Reinforcement Learning (RL) has been used in [27] to study a point-to-point wireless communication system in which the transmitter is equipped with an energy harvesting device and a rechargeable battery. The authors have shown that, the RL approach is able to reach the optimal performance of the online optimization problem, solved using policy iteration algorithm. This proposal poses the theoretical basis on the application of RL to energy harvesting networks; however, it is not applicable to multi-agent systems. RL has also been used in [28] to control a single EH SBS as a function of the local harvesting process and storage conditions. The authors design the SBS control with a Fuzzy Q-learning algorithm aimed at improving the battery lifetime and minimizing the electricity expenditures. However, the effect of the simultaneous switching OFFs by multiple SBSs on the overall network performance is not studied.

In [29], we have considered a scenario where grid powered MBSs are supported by SBSs equipped with solar panels and batteries. The solution adopted is based on distributed Q-learning. Each SBS implements an agent that, by taking autonomous decisions, independently learns switch ON/OFF policies according to the energy income and the traffic demand. This solution, despite showing encouraging results, presents problems when scaling up the number of agents due to their conflicting behaviors, since they experience difficulties in coordinating.

As a conclusion, the LL solution proposed in this paper represents the first attempt to coordinate ON/OFF switches of SBSs in a densified HetNet with renewable energy capabilities.

III. SYSTEM MODEL

A. Problem Statement

We consider a two-tier network composed of clusters of one MBS and \( N \) non-overlapping SBSs. The MBS provides baseline connectivity and is powered by the electric grid. The SBSs are deployed for increasing the capacity in a hot-spot manner (e.g., shopping malls, city centers, etc.) [30]. SBSs are solely powered through the energy harvested by a solar panel and are equipped with rechargeable batteries.

We consider a stationary discrete-time dynamic system that evolves in slots according to the variation of the traffic demand and the energy arrivals in time. The time granularity \( \Delta T \) is the time difference between two consecutive slots. The traffic load experienced in slot \( t \) by each SBS, generated by the User Equipments (UEs) in their coverage, is defined as \( L_t = [L_1^t, L_2^t, \ldots, L_N^t] \). The energy harvested by the SBSs in slot \( t \) is defined by \( E_t = [E_1^t, E_2^t, \ldots, E_N^t] \) and the energy stored by each SBS in slot \( t \) is defined by \( B_t = [B_1^t, B_2^t, \ldots, B_N^t] \). In particular, the energy stored in the batteries \( B_{t+1} \) at the beginning of the next slot is evaluated according to the following formula:

\[
B_{t+1} = \min \left( B_t + E_t - P_t \Delta T, B_{cap} \right)
\]

where \( P_t = [P_1^t, P_2^t, \ldots, P_N^t] \) is the power consumed by the SBSs in slot \( t \) and \( B_{cap} \) is the maximum battery capacity. The amount of harvested energy that exceeds the battery capacity is wasted since it cannot be stored. At each slot \( t \), the configuration of the SBSs in terms of ON/OFF states is decided based on their available energy budget and the traffic demand of the cluster. The total cost minimization problem is defined by a Markov Decision Process (MDP) as \( x_{t+1} = f(x_t, a_t, w_t) \), where \( x_t = [x_1^t, x_2^t, \ldots, x_N^t] \) is the state of the SBSs in slot \( t \), \( a_t = [a_1^t, a_2^t, \ldots, a_N^t] \) is the control, and \( w_t = [w_1^t, w_2^t, \ldots, w_N^t] \) is the random disturbance. In particular, each element \( x_i^t \), with \( i = 1, \ldots, N \), characterizes the environment and is defined as \( x_i^t = (E_i^t, B_i^t, L_i^t) \). The system control \( a_i^t \) is defined as follows:

\[
a_i^t = \begin{cases} 
0, & i\text{-th SBS is OFF} \\
1, & i\text{-th SBS is ON}
\end{cases}
\]

When a SBS is switched OFF, the associated UEs have to connect to the MBS. However, in case the MBS is reaching its capacity limit, some of the UEs may not be served and may be dropped. The probability to be in this situation, defined as system outage, must be minimized. Moreover, each battery at the SBS has to be maintained in the proper State Of Charge (SOC) to avoid a rapid reduction of its lifetime [31]. In fact, an erroneous battery utilization can permanently damage the storage capabilities and entail huge expenses for the replacement. Therefore, the optimization goal is to minimize the grid energy consumption of each cluster, while serving the whole traffic.
requests (or minimizing the losses) and maintaining the battery level of each SBS above a security threshold. The general optimization problem may be formulated as follows:

\[
P_1: \min_{\{S_i\}_{t=1,...,K}} \sum_{i=1}^{K} g(x_t, a_t, w_t) \\
\text{subject to } B_{t}^{i} > B_{th}^{OFF} \forall i.
\]

(3)

where \(B_{th}^{OFF}\) is the battery threshold level and \(K\) is the time horizon or the number of times the control is applied; \(g(x_t, a_t, w_t)\) is the weighted cost function in slot \(t\), which is defined as:

\[
g(x_t, a_t, w_t) = \omega_1 \cdot P_m(x_t, a_t, w_t) + \omega_2 \cdot D(x_t, a_t, w_t)
\]

(4)

where \(P_m(x_t, a_t, w_t)\) and \(D(x_t, a_t, w_t)\) are respectively the normalized grid power consumption and the system drop rate in the cluster, given the operative modes of the SBSs and the slot \(t\). The grid power consumption in slot \(t\) is due to the power drained by the MBS. The system drop rate in slot \(t\), \(D(x_t, a_t, w_t)\), is the ratio of total amount of traffic demand that cannot be served by the system in the time step \(t\), i.e., the sum of the traffic of all the UEs that are not served by their SBSs, which is switched OFF, nor by the MBS, which is overloaded. The weights \(\omega_1\) and \(\omega_2\) provide flexibility in the cost function to emphasize one part of the cost over the other. They must always sum to 1, that is, \(\omega_1 + \omega_2 = 1\). In this work, we will consider \(\omega_1 = \omega_2 = 0.5\) to have a balanced importance of the two components.

An offline solution of this problem is proposed in our work [32] and [33] using Dynamic Programming (DP) and with a priori knowledge of the environmental variables. In detail, the problem is represented as a graph and stated as a Shortest Path search. Label Correcting Method is used to explore the graph and find the optimal ON/OFF policy. The results achieved may be considered as system performance limits and are compared with the LL solution in this paper.

B. Control Architecture

We study the problem introduced in the above through a multi-agent system based on MRL, where each agent is located at the SBSs.

Multi-agent systems are an effective way to treat complex, large and unpredictable problems since they offer modularity in distributing the implementation of the solution across different agents. The agents can be either endowed with an offline behavior or learn new behaviors online, such that the performance of the single agent or of the whole system are improved iteratively. The offline scheme is usually solved thanks to game theory. However, sometimes the complexity of the environment makes difficult or impossible the design of the agent behavior a priori. This is the case of our densified HetNet scenario, in which the number of variables and operations is too high and makes the problem hard to model. Therefore, the online MRL solution represents a viable approach for optimizing the agents behaviors without any prior knowledge of the system dynamics. In this approach, at each slot \(t\), the agents take an action according to the state observed from their local environment. The action causes the environment to transit to a new state and allows the agents to evaluate the benefits incurred in the transition, the so called reward. By trying different actions, the agents can learn the optimal behavior of the system through the cumulative rewards. This phase is called exploration and is aimed at training the algorithm for the stable phase, called exploitation, in which the agents will use the learned policies. The exploration is a highly sensitive task since the agents explore to obtain information not only about their local environment, but also about the other agents and to adapt to their behavior. In fact, any agent’s action on the environment depends also on the action taken by the other agents. Each agent is, therefore, faced with a moving-target learning problem: the best policy changes as the other policies of agents change [34].

Our proposal of agent coordination is based on the LL paradigm [15]. The global solution is obtained in a hierarchical fashion by decomposing the multi-agent problem in different learning subtasks at each layer, anyone of which, in turn, aims at facilitating the learning process of the next higher level subtask. A heuristic function, inspired to [16], is used for interfacing the algorithms at the different layers and guiding the agents’ exploration of the state-action space in a more efficient way. In particular, the proposed LL solution is based on two-layers in charge of local and network-wide EE control, respectively. In this way, the coordination of the agents at the local layer is assisted by a centralized entity at a higher layer that has a global view of the effect of the agents’ actions on the overall environment. The goal is to help the agents in their coordination with a heuristic function that guides the exploration from a centralized and global perspective.

The first layer is a set of local agents in the SBSs, each of those in charge of learning switching ON/OFF policies according to the harvested energy arrivals, the available energy budget, the user traffic demand and the energy consumption of the SBSs. To address this objective, agents implement an algorithm based on Heuristically Accelerated MRL (HAMRL) [16]. HAMRL has been already successfully applied in the wireless communication domain in the field of Inter-Cell Interference Coordination (ICIC) problem in [35]. In this work, HAMRL has been utilized to implement a decentralized ICIC controller aimed at reducing the interference in the LTE downlink channel of a network of MBSs.

Layer 2 is a central manager, which assists the learning process based on a network-wide perspective. It is in charge of collecting local agents state information and assisting Layer 1 through the heuristic function of HAMRL. In particular, the second layer is composed of:

- A Multilayer Feedforward Neural Network (MFNN) to forecast the MBS load based on the environmental variables of each SBS.
- A SBS Centralized Controller that decides whether to enforce the local policies of a specific set of SBSs as a function of the level of congestion of the MBS.

The MFNN is used to estimate the load of the MBS only with partial knowledge of the amount of resources needed by the users to be served and saves overhead signalling. In fact, traffic
load at the MBS depends on the position of the users within the cell, which differs in case they are served by a SBS or the MBS. As a result, the agents at each SBS in Layer 1 can keep a local view of the system and avoid large state spaces for maintaining the algorithm complexity at reasonable levels. This also prevents from having too large phase of exploration, which, in turn, reduces the possibility of generating conflicts among the agents. The algorithms implemented in the two layers will be detailed in Sections IV and V.

The two layers interact thanks to a SDN framework that provides the infrastructure for collecting the needed parameters and distributing the control policies. Fig. 1 shows the architecture proposed in this paper. An example of such SDN solution is the Energy Monitoring & Management Application (EMMA) defined in the 5G-Crosshaul project [36]. EMMA is an infrastructure-related application based on the SDN paradigm aimed at monitoring the status of the RAN, fronthaul and backhaul elements and triggering reactions to minimize the energy footprint.

C. Energy Consumption and Harvesting Models

The BS power consumption model adopted is $P = P_0 + \beta \rho$, where $\rho \in [0, 1]$ is the BS traffic load, normalized with respect to its maximum capacity, and $P_0$ is its baseline power consumption. This model is supported by real measurements [37] and closely matches the real power profile of LTE BSs.

The energy harvesting traces are generated based on a Markov model that provides accurate statistics per month basis by processing the hourly solar energy arrival data over 20 years [9] and used for simulation purposes. In detail, the 24 hours are divided into a number $N_s \geq 2$ of time slots of constant duration, equal to $T_h$ hours, $i = 1, \ldots, N_s$. Each slot is a state of the Markov model and the pdf is evaluated through the kernel smoothing technique per month basis, considering the empirical data that has been measured for all days in the dataset for the month under consideration.

D. Traffic Model

The UE resource allocation scheme uses the methodology defined in [38]. This includes a detailed wireless channel model and the dynamic selection of the Modulation and Coding Scheme (MCS) for each UE as function of its Signal to Interference plus Noise Ratio (SINR), which is given by:

$$\text{SINR} = \frac{|h_0|^2 P_{t,0}}{\sum_{i=1}^{N_I} |h_i|^2 P_{t,i} + \sigma_0^2},$$

where $P_{t,0}$ and $h_0$ are the transmission power and the channel gain for the useful transmission respectively, $N_I$ is the number of interferers, whereas $|h_i|^2$ and $P_{t,i}$ represent the channel gain and the transmission power of the $i$-th interferer. $\sigma_0^2$ is the power of the thermal noise. The UEs have been classified as in [37] in heavy and ordinary users according to their amount of requested traffic. The traffic demand of each UE in a cycle has been dimensioned for mimicking the realistic traffic profiles presented in [21], which are derived combining time, location and frequency information of thousands of cellular towers. The analysis in [21] demonstrates that the urban mobile traffic usage can be described by only five basic time domain patterns that corresponds to functional regions, i.e., residential, office, transportation, entertainment and comprehensive. We will consider the first two regions (residential and office), which represent the most frequent patterns in the studied area.

IV. LAYER 1: LOCAL OPTIMIZATION

A. Distributed Q-Learning and HAMRL Overview

The first layer is composed of a set of distributed agents implementing HAMRL with the goal of dynamically switching ON and OFF the SBSs according to the available harvested energy budget, the user traffic demand and the energy consumption of the SBS. The control decisions are made by multiple intelligent and uncoordinated agents, which can only partially observe the overall scenario. Therefore, the local environment may differ from agent to agent, since they come from spatially distributed sources of information. To this end, the HAMRL extension of the distributed Q-learning [39] technique has been considered in this paper, and briefly introduced next.

The decision making problem in distributed Q-learning is based on the MDP defined in Section III-A. In detail, each SBS implement an agent $i$ in charge of maintaining a local policy and a local Q-function $Q(x^i_t, a^i_t)$ representing the level of convenience in taking actions $a^i_t$ in state $x^i_t$, with $r$ being the decision epoch (slot). As a result of the execution of this action, the environment returns an agent dependent reward $r^i_t$, which allows the local update of a Q-value, $Q(x^i_t, a^i_t)$. The Q-value is computed according to the rule:

$$Q(x^i_t, a^i_t) \leftarrow Q(x^i_t, a^i_t) + \alpha \left[ r^i_t + \gamma \max_{a'} Q(x^i_{t+1}, a') - Q(x^i_t, a^i_t) \right]$$

Fig. 1. Layered Learning control architecture overview.
where $\alpha$ is the learning rate, $\gamma$ is the discount factor, $x_{i+1}^t$ is the next state for agent $i$ and $a'$ is the associated optimal action. For more details on RL and Q-learning the reader is referred to, e.g., [39]. Consequently, thanks to Eq. (6), the control policy can be learned by exploring the environment. The most common exploration procedure in Q-learning is the $\varepsilon$-greedy [39], which consists in having a learning agent policy $\Pi(x_i^t)$ that randomly chooses a sub-optimal action with probability $1-\varepsilon$, i.e., $\Pi(x_i^t) = \arg \max_{a_i} (Q(x_i^t, a_i^t))$. The choice of the maximum Q-value makes Q-learning an off-policy algorithm, since the policy to generate the behavior is unrelated to the policy evaluated. In fact, it uses the greedy exploration for estimating the long term reward, while the Q-values are evaluated according to Eq. (6).

Alternatively, HAMRL guides the exploration by introducing a heuristic function $H(x_i^t, a_i^t)$ in the learning agent policy $\Pi(x_i^t)$ for influencing its action choices. In particular, $H(x_i^t, a_i^t)$ contains additional knowledge not included in the state variables (Layer 2 in our case). The resulting HAMRL policy selection formula is:

$$\Pi(x_i^t) = \arg \max_{a_i}
(Q(x_i^t, a_i^t) + H(x_i^t, a_i^t))$$

(7)

$H(x_i^t, a_i^t)$ has to assume values and dimensions comparable with those in the agent Q-tables to properly influence the action to be taken and the policy $\Pi(x_i^t)$. If $H(x_i^t, a_i^t) = 0$, the algorithm behaves like a regular Q-learning for the $i$-th SBS.

**B. HAMRL for Energy Harvesting SBSs**

We define the local state $x_i^t$ of agent $i$ in slot $t$ as $x_i^t = (E_i^t, B_i^t, L_i^t)$. The three variables characterizing the environment represent the instantaneous energy harvested, the battery level and the SBS load, respectively. Since these parameters assume continuous values, they have been normalized and quantized for having a reasonable number of states to be explored, in detail $E_i^t$, $B_i^t$ and $L_i^t$ into 2, 5 and 3 levels, respectively. The possible actions are to switching ON and OFF the SBS.

The reward of an agent $i$ in slot $t$ is defined as follows:

$$r_i^t = \begin{cases} 0 & B_i^t < B_{th}^{OFF} \text{ or } D_t > D_{th} \\ \kappa T_i^t & B_i^t \geq B_{th}^{OFF} \text{ and } D_t \leq D_{th} \text{ and SBS } i \text{ is ON} \\ 1/B_i^t & B_i^t \geq B_{th}^{OFF} \text{ and } D_t \leq D_{th} \text{ and SBS } i \text{ is OFF} \\ \end{cases}$$

(8)

where $T_i^t$ is the normalized throughput of SBS $i$ in slot $t$, $D_t$ is the instantaneous system drop rate, defined as the ratio between the total amount of traffic dropped and the traffic demand in the entire cluster. $D_{th}$ is a design parameter that represents the maximum tolerable system drop rate. Finally, $B_{th}^{OFF}$ is the threshold on the battery level needed to maintain it in the proper SOC.

The rationale behind Eq. (8) is to mimic the optimization problem in Eq. (3). The condition in the first line implies a zero reward when the battery level falls below $B_{th}^{OFF}$ ($B_i^t < B_{th}^{OFF}$) or the system drop rate is below $D_{th}$ ($D_t < D_{th}$), and aims at penalizing the action that has been taken since it has led to a critical condition. This incentives the SBS to turn itself OFF to save energy, as this implies a higher reward. Thus, the SBS can be switched ON and offload the macro BS later. Whereas, when $D_t > D_{th}$, the system performance is deemed not sufficient and the SBS is encouraged to stay ON and support the MBS to serve the traffic requests. In the second and third line of Eq. (8), the reward is proportional to the throughput when the SBS is turned ON and is, instead, proportional to the inverse of the energy buffer level when the SBS is OFF. Note that the SBS, after a training phase, will choose to remain ON (and offload the MBS) when the reward in the second line is higher, i.e., when $\kappa T_i^t > 1/B_i^t$. Note that $1/B_i^t$ may dominate over $\kappa T_i^t$ in case battery level and throughput are both low. In this case, the SBS switches OFF to save energy. The constant $\kappa$ is used to balance the impact of the two terms in the reward function.

**V. LAYER 2: CENTRAL MANAGER**

The task of this layer is of guiding the agents of the Layer 1 toward more EE switch ON/OFF policies. To achieve this, we advocate for using a network-wide view to reduce the system traffic drop rate and to concurrently manage more efficiently the harvested energy. In particular, the second layer is in charge of deciding the local agent(s) to be influenced and returns the most appropriate set of heuristics values $H_t = [H(x_1^t, a_1^t), H(x_2^t, a_2^t), \ldots, H(x_N^t, a_N^t)]$ according to network-wide parameters that are not present in the HAMRL optimization. Layer 2 comprises an MBS load estimator based on a MFNN which provides the input to a SBS Centralized Controller, in charge of evaluating the system load conditions of the whole cluster and selecting the SBSs to be influenced.

**A. MBS Load Estimator**

A MFNN estimates the normalized MBS load $L_{MBS}^t$ in the slot $t$ as a function of the load of each SBS $L_i^t$ and of their ON/OFF policy $\Pi_{t} = [\Pi(x_1^t, a_1^t), \Pi(x_2^t, a_2^t), \ldots, \Pi(x_N^t, a_N^t)]$. We define the estimated load of the MBS in the slot $t$ as $L_{MBS}^t$. A supervised approach has been adopted, i.e., a training set of input-output is used to train the MFNN according to the backpropagation algorithm [40].

The basic element of a MFNN is represented by the neuron (also called perceptron), which consists of a linear combination of fixed non-linear functions $\theta_j(x)$. In detail, for a vector of input $x_i, i = 1, \ldots, N$, it takes the form:

$$y(x, \vartheta) = p \left( \sum_{j=1}^{N} \vartheta_j \theta_j(x) \right)$$

(9)

where $\vartheta_i$ are the weights associated to each input and $p(\cdot)$ is a non-linear activation function, typically the sigmoid function $p(x) = 1/(1 + e^{-x})$. A MFNN is composed of a series of neurons organized in $M$ layers in a way that the input information moves only in one direction (i.e., there are no cycle in the networks like in a recurrent ANN). Let define $I$ as the number of neurons in layer $m$. The bottom layer, $M_0$, is the input layer and it contains $N + 1$ neurons, which are the inputs plus the constant neuron always at 1. The last layer is
composed by only one neuron and represents the output of the ANN. Each neuron in a layer \( m = 2, \ldots, M \) has \( I_{m-1} \) inputs, each of which is connected to the output of a neuron in the previous layer. Layers \( 2, \ldots, M-1 \) are called hidden layers. A MFNN can approximate arbitrary continuous functions defined over compact subsets of \( \mathbb{R}^N \) by using a sufficient number of neurons at the hidden layers. In order to achieve this, it is necessary to determine the values of the weights correspondent to the function to be approximated (also known as training phase). Further details will be provided in Section VI-B.

### B. SBS Centralized Controller

We identify two different critical cases based on the load of the MBS:

1. The system is under-dimensioned when \( \hat{L}_t^t \) MBS is above the threshold \( L_{\text{thrHigh}}^t \).
2. The system is over-dimensioned when \( \hat{L}_t^t \) MBS is below the threshold \( L_{\text{thrLow}}^t \).

The first case occurs when many SBSs are OFF simultaneously. In this case, the switching OFF of some SBSs can be delayed, whether their battery levels allow such operation, to increase the system throughput. In more detail, the SBS Centralized Controller defines the set of candidate SBSs that can be switched ON (\( SBS_{\text{ON}}^t \)) among those that are in OFF state and have enough battery reserves. Alternatively, in the second case, the traffic demand can be served by a lower number of SBSs and be managed by the MBS with negligible impact in the grid energy consumption. In detail, the set of candidate SBSs to be switched OFF (\( SBS_{\text{OFF}}^t \)) is defined among those that are in ON state and have scarce energy reserves (i.e., \( B_{t_i}^t \leq B_{\text{LOW}}^t \)).

The number of SBSs to be influenced and their relevant heuristics values \( H_i^t \) are derived based on Algorithm 1. For each SBS \( i \) that has been activated or deactivated in this process, the correspondent heuristic value is set to \( -Q_{1,\text{MAX}}^t = -\max_{a_i^t} (Q(x_i^t, a_i^t)) \) in \( H_i^t \). In fact, \( H(x_i^t, a_i^t) \) must be the lowest value that can influence the action in order to minimize the distortion in the Q-value function due to the use of heuristics [16]. Therefore, for influencing the choice of action of SBS \( i \) in state \( x_i^t \), \( H(x_i^t, a_i^t) \) should be negative and higher than the maximum Q-value in \( x_i^t \) (i.e., \( Q_{1,\text{MAX}}^t \)).

---

**VI. NUMERICAL RESULTS AND DISCUSSION**

### A. Simulation Scenario

The scenario considered in this analysis is composed of a single cluster with 1 MBS placed in the middle of a \( 1 \times 1 \) km\(^2\) area and a varying number of SBSs randomly placed and non-overlapping. We consider medium scale factor “metro cells” as SBSs, featuring a maximum transmission power of 38 dBm, which corresponds approximately to 50 meters of coverage range. The values of \( \beta \) and \( P_0 \) of the energy model presented in Section III-C for the MBS (SBS) are 600 (39)W and 750 (105.6), respectively.

Each SBS is supplied by an array of 16 \( \times \) 16 solar cells of Panasonic N235B solar modules (area 4.48 m\(^2\)), that have single cell efficiencies of about 21%, and a lithium ion battery of 1.5 KWh. This dimensioning of the RES equipment has been proved to be optimal for the winter season through simulations using the algorithm in [32]. The solar energy arrivals are generated according to Section III-C for the city of Los Angeles considering \( N = 12 \) states. The traffic demand is modeled as in Section III-D. In detail, the office and residential traffic profiles have been considered, respectively termed “Res” and “Off” in the figures. Both of them present a high activity during the day. However, they differ in the profile: the traffic is concentrated during the daylight hours (e.g., from 10 am to 6 pm) in office, while the residential has only one peak during the early night hours (e.g., from 6 pm to 12 pm). User traffic has been classified in two classes, namely heavy and ordinary, according to [37]. The main simulation parameters are given in Table I.

### B. MFNN Training Analysis

We analyze here the behavior of the MFNN in Layer 2 during the training phase. Based on our simulative analysis, the best number of neurons per layer is \( I_1 = \left\lceil 3 N / 2 \right\rceil \), \( I_2 = \left\lceil 2 N / 3 \right\rceil \) and \( I_3 = \left\lceil 2 N / 3 \right\rceil \). Fig. 2 presents the overall mean squared error (mse) of this configuration for MFNN with two and three hidden layers (“2L” and “3L” in what follows, respectively) as a function of the day, which includes 24 system evolution slots. The tests have been carried out for a cluster of 10 SBSs, which represents a complex scenario, as the approximated function increases its domain \( R^N \) in dimension. The 2L starts with lower mse and performs better already at 50 days, reaching mse values below 0.2. Alternatively, the
### TABLE I

**SIMULATION PARAMETERS**

| Parameter                        | Value         |
|----------------------------------|---------------|
| Scenario                         | Solar panel size (m²) | 4.48 (16×16) |
|                                  | Solar panel efficiency (%) | 21           |
|                                  | Battery capacity (kWh)   | 1.5          |
|                                  | MBS transmission power (dBm) | 43          |
|                                  | SBS transmission power (dBm) | 38          |
|                                  | Bandwidth (MHz)         | 5            |
|                                  | Heavy users traffic (MB/h) | 900         |
|                                  | Ordinary users traffic (MB/h) | 112.5      |
|                                  | Epoch duration (h)       | 1            |
| HAMRL                            | α              | 0.5          |
|                                  | γ              | 0.9          |
|                                  | ε              | 0.1          |
|                                  | κ              | 10           |
|                                  | Battery threshold $B_{\text{thr} MBS}^{\text{OFF}}$ (%) | 20           |
|                                  | System drop rate threshold $D_{\text{th}}$ (%) | 3            |
| MFNN                             | Learning rate (%)     | 0.1          |
|                                  | Activation function   | $f(x) = 1/(1 + e^{-x})$ |
|                                  | First hidden layer no. | $I_1 = [N/3/2]$ |
|                                  | Second hidden layer no. | $I_2 = [N/2/3]$ |
|                                  | Third hidden layer no. (when present) | $I_3 = [N/2/3]$ |
| SBS Centralized Controller      | High congestion threshold $L_{MBS}^{\text{THIGH}}$ (%) | 85           |
|                                  | Low load threshold $L_{MBS}^{\text{LOW}}$ (%) | 5            |
|                                  | Scarse battery threshold $B_{\text{thr} MBS}^{\text{LOW}}$ (%) | 20           |

**Fig. 2.** Mean squared error of the MFNN for different number of hidden layers.

3L presents higher mse till 100 days; after that, its mse starts decreasing faster and reaches lower values asymptotically.

As an additional illustrative result, we evaluate the statistical measure called sensitivity, which is defined as the proportion of positive cases that are correctly identified as such, in detail:

\[
\text{sensitivity} = \frac{\text{true positive no.}}{\text{true positive no.} + \text{false negative no.}} \tag{10}
\]

where we define the false negatives as the cases when the MFNN does not estimate that the system is under-dimensioned (i.e., $\hat{L}_{MBS}^t \leq L_{MBS}^{\text{thrHigh}}$) but it is really in outage. Fig. 3a provides the sensitivity as a function of the day. From Fig. 3a, we can observe that the 2L takes approximately 50 days for reaching a stable behavior, whereas the 3L takes 10 times longer and passes the 500 days. Besides, the specificity measures the proportion of negative cases that are correctly identified as such, which corresponds to:

\[
\text{specificity} = \frac{\text{true negative no.}}{\text{true negative no.} + \text{false positive no.}} \tag{11}
\]

where false positives have been defined as the cases when the MFNN expects that the system is under-dimensioned (i.e., $\hat{L}_{MBS}^t > L_{MBS}^{\text{thrHigh}}$) but it is not in outage. Fig. 3b depicts the specificity as a function of the day. In this case, the MFNNs reach a stable behavior at 1500 days. However, the 2L presents less variance on the specificity. We can also note that the asymptotic value of the specificity is lower than the sensitivity. This is due to the fact that we have adopted a guard margin to avoid the MBS to be overloaded (i.e., $L_{MBS}^{\text{thrHigh}} = 0.85$). Therefore, some false positives are MFNN estimations that fall between $L_{MBS}^{\text{thrHigh}}$ and 1, which do not represent a system outage.

Based on the analysis in the above, a MFNN with two hidden layers has been used in the analysis in the following subsections due to its better sensitivity and specificity, and a faster training phase.

### C. Layered Learning Training Analysis

The training phase of the LL framework has been evaluated compared with a distributed Q-learning solution based on our proposal in [29], “QL” in the following. QL is a standard MRL solution defined with the same environmental model described in Section IV-B, which uses the standard Q-learning policy selection formula (i.e., without the heuristic function). The training has been analyzed considering the stability of the system when starting with all the Q-values initialized to zero to avoid conditions of battery degradation, which can occur when battery level drops below the security SOC threshold $B_{\text{OFF}}$.

An example of the behavior of LL and QL algorithms is shown in Fig. 4 and Fig. 5, where the hourly battery level of...
a SBS is plotted for a scenario with 3 SBSs and different traffic profiles. The simulations start with the month of January, and runs for 400 days, spanning across the correspondent months. In both cases, the system starts with a short-sighted approach, since it is using the energy only according to the instantaneous availability, and drops frequently below the threshold $B_{th}^{OFF}$. In fact, during this period, the agent is at the beginning of the exploration phase and has to gather information from the environment to update its Q-values accordingly. The training phase of the office traffic profile results to be the shortest, with about 40 days for both 20% and 50% of heavy users. With residential traffic profile, the LL takes 50 and 80 days for reaching a stable behavior for the case of 20% and 50% of heavy users, respectively. After these points, the battery level drops less often below $B_{th}^{OFF}$ and the density of points starts becoming more prominent above the battery threshold. Similarly to what was experienced with the duration of the training phase, the number of points falling below the threshold in case of office traffic profile is the slightest. This phenomenon is due to the fact that the traffic in the office scenario is concentrated during the daylight hours, as for the energy in the harvesting process, which implies that SBSs have more energy when it is needed. Alternatively, the residential traffic profile has the peak during the early night hours (e.g., from 8 PM to 12 PM), which is when the solar energy scavenged starts being scarce. In such a case, the improvement of the LL approach with respect to the QL is more evident, as depicted in Fig. 5a and Fig. 5b. In fact, in Fig. 5a the LL is able to avoid that the battery level falls outside the ideal SOC window (i.e., below $B_{th}^{OFF}$), while QL presents many points below the battery security threshold starting from 300 days, which is approximately the beginning of the winter season. The effect of the traffic demand can be appreciated in both Fig. 4 and Fig. 5. In case of the office traffic profile, the minimum average battery level decreases from 0.6 to 0.4. In case of the residential traffic profile, only the LL is able to guarantee the minimum battery level and only for the case of 20% of heavy users. It is to be noted that, the energy harvesting equipment is dimensioned for the worst case scenario of working in the winter season. Thus, during the summer the energy reserves are abundant and, consequently, the learning processes are facilitated in finding a solution for maintaining the battery level above the $B_{th}^{OFF}$ threshold.
Finally, it is to be noted that, even after the training phase has finished, a good trade-off between exploration and exploitation has been adopted in order to guarantee a constant update of the Q-table and the selection of the proper ON/OFF policy across the different seasons, where the energy harvesting process can have substantial variations. These changes in the environment are well managed by the learning algorithm as presented in [29], where we proved that the difference in network performance between QL and an offline trained QL are negligible. In detail, the offline QL uses pre-calculated Q-tables per month basis to avoid their update due to the variations of the energy process across the months.

D. ON/OFF Policies

In this section, we analyze the behavior of the switch ON/OFF policies of the LL solution. The policies obtained offline with a Direct Load Control based on DP presented in Section III-A, “Offline” in the following, are presented for sake of comparison. The policies of LL have been evaluated across 365 days of simulation with the training already performed offline. The results are presented separately for the winter and the summer periods, respectively termed “Win” and “Sum” in the plots, since the harvesting process substantially differs for different seasons. January, February, October, November and December are considered winter months. The network scenario studied is based on a cluster of 3 SBSs.

Fig. 6 shows that the policies substantially converge in having a high switch OFF rate during the night to save energy for the daily peak of traffic. However, the LL algorithm is starting to highly switch OFF earlier, at 8 pm and 12 pm for 20% and 50% of heavy users, respectively. As it can be noted, the
Fig. 7. Daily average switch OFF rate for the LL and Offline solutions with residential traffic profile. Scenario with 70 UEs per SBS with 20% and 50% of heavy users.

The total amount of traffic in the network influences the policies of the LL algorithm reducing the high switch OFF rate window duration from 10 to 6 hours. Therefore, the main difference between the Offline and the LL solutions is in the extent of the high switch OFF rate window, where the LL presents a more conservative approach and needs to switch OFF with higher intensity for being able to reach the design goals. This phenomenon is mainly due to the random nature of the energy harvesting process, which may provide very diverse energy incomes across different days. Fig. 7 shows that the policies of the two schemes have a similar behavior during the night and differs during the day. In fact, considering the case of 50% of heavy users in Fig. 7b, the Offline solution reports an extra switching OFF peak during the afternoon in winter, from 2 to 3 pm, in order to save energy for the peak of traffic during the night. On the contrary, LL is maintaining the behavior of the case of 20% of heavy users with only switch OFF window during the night. However, LL reacts to the higher traffic demand during the night by reducing of 50% the switch OFF rate window duration with respect to the case of the office traffic profile. It is to be noted that, this is the hardest case for LL stability, as testified by Fig. 5b.

E. Network Performance

In this section the LL framework is compared against the QL solution introduced in Section VI-C and a greedy (GR) algorithm. The GR switches OFF a SBS when its battery is below a security threshold $B_{th}^{OFF}$, and reactivates it when the battery returns above the threshold. Results are obtained averaging simulations spanning over different months for an overall duration of 365 simulated days with LL and QL already trained. It is worth noting that, the performance behavior evaluated including the training phase does not change substantially, since the training phase is relatively short with respect to the assessment window, as presented in Section VI-C, and will not be reported in the following. Despite of the fact that the training is performed offline, the exploration phase is not stopped to be able to follow the slower dynamics of the harvesting energy process across the seasons. As for Section VI-D, the results will be presented analyzing the two representative periods “Win” and “Sum”. We considered a high-traffic scenario involving 70 UEs with 50% of heavy users at each SBSs. The results with low-traffic presents a similar behavior and will not be presented due to space reasons.

Fig. 8 presents the system average percentage gain in throughput of the LL and QL schemes with respect to the GR. The LL framework presents always a higher throughput. Moreover, the LL throughput gain is increasing with the number of SBSs, while with QL the gain starts reducing from 6 SBSs. This phenomenon is of particular intensity in case of residential traffic profile, where QL has a throughput lower than the GR in the summer period, as depicted in Fig. 8a. This is originated by the typical problem of MRL, discussed in Sections I and III, where the lack of coordination may generate conflicting behaviors among the agents. This issue may occur with higher probability when the number of agents grows, as clearly demonstrated by the QL performance in Fig. 8. Alternatively, LL is able to improve the coordination of the agents and better exploits the increased capacity resources. It is to be noted that, during summer the gain in throughput is lower since the renewable energy system has been dimensioned to provide the necessary energy in winter season. This implies that the harvested energy in summer is generous and both LL and QL have fewer margins for policy optimization.

Fig. 9 reports the average system drop rate of the three schemes, as defined in Section III-A. LL always presents the lowest system drop rate and maintains it, almost always, below the threshold $D_{th}$ of the 3%. The only exception is for the case of residential traffic profile in the winter season, where the system drop rate reaches the 8%. Alternatively, the QL solution presents a higher system drop rate with respect to the GR in several cases. For residential traffic profile, GR outperforms QL with 10 SBSs during the summer, and for the office traffic profile both in summer, starting from 8 SBSs, and in winter, in case of 10 SBSs. This confirms the problem of QL when scaling up the number of SBSs.
We now analyze the average daily performance during the summer and winter periods in the scenario with 10 SBSs to highlight the differences between QL and LL in the most sensitive hours. In Fig. 10, we report the system drop rate of the LL, QL and GR solutions in a cluster of 10 SBSs and varying the number of UEs per SBS. The LL solution is able
to reduce the system drop rate of more than 50% with respect to GR. On the contrary, QL has always the worst drop rate both in summer and winter period starting from 60 UEs per SBS and considering the office traffic profile. Finally, Fig. 11 presents the average hourly system drop rate for the case of 70 UEs per SBS. It is clear that LL outperforms the other solutions during all the day for both traffic profiles. In detail, in case of the office traffic profile, LL is able to stay below 3%, $D_{th}$, during the whole day, while GR and QL present high peaks in the early morning (25% at 9 am) and early night (15% from 8 pm to 12 pm). Regarding the residential traffic profile, LL passes the threshold $D_{th}$, since it experiences difficulties in managing the high traffic peaks at late night and early morning, which are two sensitive periods as, in both of them, the system has very limited energy reserves.

F. Energy Assessment

Table II presents the footprint of the two learning-based methods (LL and QL) together with a baseline solution where both the MBS and the SBSs are powered with the grid. In particular, the comparison is performed in terms of grid energy consumption for a scenario with 50% of heavy users. In addition, the column named “excess energy” reports the values of the harvested energy that cannot be used by the SBSs nor stored in the batteries, since the harvesting/storage system is dimensioned for the worst case of the winter season.

The LL solution can reach energy savings of up to 50% during the summer, as for the scenarios with 10 SBSs. The savings are affected by the number of SBSs deployed. In fact, for a smaller numbers of SBSs (e.g., 5), the energy savings are of the 20−30%. The traffic profile is another important factor that influences the footprint, since it varies in the total amount of data exchanged in the network and in temporal dynamics, as discussed in Section VI-D. The energy savings for the scenarios with office traffic profile are in general 10% higher than the residential traffic profile. The reason behind this fact is that the residential traffic profile has a peak of traffic during the night (12 am), which is where the energy reserves are scarce and the learning solutions need to rely more on the MBS, as can be noticed with the longer high switch OFF rate window depicted in Fig. 7. Considering the two learning methods, the amount of traffic delivered influences the behavior of the energy consumption, as expected. In this way, LL consumes more energy since it drops less traffic. However, the gap between LL and QL is almost null when considering scenarios with 10 SBSs, which is where QL experiences the highest drop error rate, as presented in Fig. 9.

Finally, looking at the excess energy values in Table II, we can appreciate how the harvested energy process is abundant in summer, where the energy that cannot be used is considerably
higher. As for the energy savings, this circumstance occurs with particularly intensity with the residential traffic profile, that is the scenario in which the learning algorithms experience the worst network performance. The behaviors of QL and LL are similar, except for the case of summer with 10 SBSs, where the QL consumes 10% more of harvested energy, despite of the fact that is dropping more traffic. This behavior confirms that LL performs better when scaling up the number of SBSs and is able to use more efficiently the available energy reserves, since it utilizes less solar energy than QL and, concurrently, delivers more traffic.

In the light of the above results, future work needs to consider further investigation on the decisions making process for enabling the system to reduce the excess energy and work with the same performance under different traffic conditions. Moreover, considering the abundance of the excess energy, it can be taken into consideration in the control framework either for sharing it with other network elements, and/or trading it, as done by the prosumers in the smart grid architecture.

VII. CONCLUSION

In this paper we have presented a framework for the traffic control of a densified HetNet powered with solar power based on SDN/NFV and MRL. In particular, the coordination problem of finding simultaneously a solution among all the agents that is good for the whole system has been solved through Layered Learning. The first layer implements a HAMRL algorithm to learn the control policies locally at each SBS and to adapt to the dynamic conditions of the environment, in terms of energy inflow and traffic demand. Layer 2 is in charge of improving the Layer 1 policies through a MFNN-based solution by considering network-wide parameters and mitigating the effects of the conflicting behavior of the agents.

We have analyzed the training phase of the LL algorithm, which presents improvements compared to a distributed MRL solution thanks to the introduction of the layerized coordination. Then, we compared the policies with respect to an offline solution based on DP, obtaining that they act very similarly. The LL algorithm has been also contrasted against a greedy and a distributed MRL solutions from a network performance perspective. Simulations results show that the proposed solution outperforms the others in terms of throughput and drop rate. The energy savings achieved are considerable with respect to a baseline solution where both MBS and SBSs are powered with the grid, reaching up to the 50%. As a result, we can conclude that our proposal succeed in incorporating stability of the learning process on the one hand, and adaptation to the dynamic behavior of the agents on the other hand, that was our primary goal.

There are several ways in which this work can be extended. The traffic model used could incorporate accurate geographical and spatial information. In fact, densified HetNet scenarios are characterized for specific spatial distribution of users (i.e., hotspot coverage) that varies during the day according to the zone (e.g., residential, office, transportation). The proposed solution has shown some stability problems for network of very dense SBSs and high traffic. The LL scheme adopted represents a good starting point, since it combines the flexibility of a distributed learning solution with the efficiency of a centralized approach. Further work can be done for a deeper integration between the solutions adopted in the two layers, e.g., using deep learning methods. Finally, the results on the excess energy encourages to integrate its control in the optimization problem. In fact, the presence of surplus energy perfectly matches with Demand Response method of the smart grid, where the energy can be shared among elements and/or traded with the operators.

REFERENCES

[1] A. Fehske, J. Malmodin, G. Biczók, and G. Fettweis, “The global footprint of mobile communications: The ecological and economic perspective,” IEEE Commun. Mag., vol. 49, no. 8, pp. 55–62, Aug. 2011.
[2] M. P. Mills, The Cloud Begins With Cool: Big Data, Big Networks, Big Infrastructure, and Big Power, Digit. Power Group, Fremont, CA, USA, Aug. 2013.
[3] A. S. G. Andrae and T. Edler, “On global electricity usage of communication technology: Trends to 2030,” Challenges, vol. 6, no. 1, p. 117, 2015.
[4] Environmental Engineering (EE): Assessment of Mobile Network Energy Efficiency (V1.2.1), ETSI Standard ES 203 208, 2017.
[5] “Study on energy efficiency aspects of 3GPP standards (rel.14),” 3GPP, Sophia Antipolis, France, Rep. TR 21.866, 2017.
[6] F. Han, S. Zhao, L. Zhang, and J. Wu, “Survey of strategies for switching off base stations in heterogeneous networks for greener 5G systems,” IEEE Access, vol. 4, pp. 4959–4973, 2016.
[7] H. A. H. Hassan, L. Nuaymi, and A. Pelov, “Renewable energy in cellular networks: A survey,” in Proc. IEEE Online Conf. Green Commun. (GreenCom), Oct. 2013, pp. 1–7.
[8] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, “Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments,” IEEE Commun. Surveys Tuts., vol. 17, no. 4, pp. 2078–2101, 4th Quart., 2015.
[9] M. Miozzo, D. Zordan, P. Dini, and M. Rossi, “SolarStat: Modeling photovoltaic sources through stochastic Markov processes,” in Proc. IEEE Energy Conf. (ENERGYCON), Dubrovnik, Croatia, May 2014, pp. 688–695.
[10] A. G. Tsikalakis and N. D. Hatziargyriou, “Centralized control for optimizing microgrids operation,” in Proc. IEEE Power Energy Soc. Gen. Meeting, Jul. 2011, pp. 1–8.
[11] “Software-defined networking: The new norm for networks,” Menlo Park, CA, USA, Open Netw. Found., White Paper, Apr. 2016.
[12] Y. Li and M. Chen, “Software-defined network function virtualization: A survey,” IEEE Access, vol. 3, pp. 2542–2553, 2015.
[13] M. Y. Arslan, K. Sundaresan, and S. Rangarajan, “Software-defined networking in cellular radio access networks: Potential and challenges,” IEEE Commun. Mag., vol. 53, no. 1, pp. 150–156, Jan. 2015.
[14] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí, “Knowledge-based 5G radio access network planning and optimization,” in Proc. Int. Symp. Wireless Commun. Syst. (ISWCS), Sep. 2016, pp. 359–365.
[15] P. Stone and M. M. Veloso, “Layered learning,” in Proc. 11th Eur. Conf. Mach. Learn. (ECML), London, U.K., 2000, pp. 369–381. [Online]. Available: http://dl.acm.org/citation.cfm?id=645327.649544
[16] R. A. C. Bianchi, M. F. Martins, C. H. C. Ribeiro, and A. H. R. Costa, “Heuristically-accelerated multiagent reinforcement learning,” IEEE Trans. Cybern., vol. 44, no. 2, pp. 252–265, Feb. 2014.
[17] M. Miozzo and P. Dini, “Layered learning radio resource management for energy harvesting small base stations,” in Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring), Jun. 2018, pp. 1–6.
[18] G. Pinto et al., “HetNets powered by renewable energy sources,” IEEE Internet Comput., vol. 17, no. 1, pp. 32–39, Jan./Feb. 2013.
[19] D. Zordan, M. Miozzo, P. Dini, and M. Rossi, “When telecommunications networks meet energy grids: Cellular networks with energy harvesting and trading capabilities,” IEEE Commun. Mag., vol. 53, no. 6, pp. 117–123, Jun. 2015.
[20] H. S. Dhillon, Y. Li, P. Nuggehalli, Z. Pi, and J. G. Andrews, “Fundamentals of heterogeneous cellular networks with energy harvesting,” IEEE Trans. Wireless Commun., vol. 13, no. 5, pp. 2782–2797, May 2014.
[21] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, “Understanding mobile traffic patterns of large scale cellular towers in urban environment,” IEEE/ACM Trans. Netw., vol. 25, no. 2, pp. 1147–1161, Apr. 2017. [Online]. Available: https://doi.org/10.1109/TNET.2016.2623950

[22] M. A. Marsan, G. Bucalo, A. Di Caro, M. Meo, and Y. Zhang, “Towards zero grid electricity networking: Powering BSs with renewable energy sources,” in Proc. IEEE Int. Conf. Commun. (ICC), Budapest, Hungary, Jun. 2013, pp. 596–601.

[23] National Renewable Energy Laboratory (NREL). PVWatts Simulator. Accessed: Sep. 2019. [Online]. Available: https://pvwatts.nrel.gov

[24] G. Lee, W. Saad, M. Bennis, A. Mehbodniya, and F. Adachi, “Online SKI rental for ON/OFF scheduling of energy harvesting base stations,” IEEE Trans. Wireless Commun., vol. 16, no. 5, pp. 2976–2990, May 2017.

[25] A. H. Arani, M. J. Omidi, A. Mehbodniya, and F. Adachi, “Minimizing base stations’ ON/OFF switchings in self-organizing heterogeneous networks: A distributed satisfactory framework,” IEEE Access, vol. 5, pp. 26267–26278, 2017.

[26] A. P. C. da Silva, D. Renga, M. Meo, and M. A. Marsan, “The impact of quantization on the design of solar power systems for cellular base stations,” IEEE Trans. Green Commun. Netw., vol. 2, no. 1, pp. 260–274, Mar. 2018.

[27] P. Blasco, D. Gunduz, and M. Dohler, “A learning theoretic approach to energy harvesting communication system optimization; IEEE Trans. Wireless Commun., vol. 12, no. 4, pp. 1872–1882, Apr. 2013.

[28] M. Mendil, A. D. Domenico, V. Heiries, R. Caire, and N. Hadsaid, “Battery-aware optimization of green small cells: Sizing and energy management,” IEEE Trans. Green Commun. Netw., vol. 2, no. 3, pp. 635–651, Sep. 2018.

[29] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, “Switch-ON/OFF policies for energy harvesting small cells through distributed Q-learning,” in Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW), Mar. 2017, pp. 1–6.

[30] NGMN KPIs and Deployment Scenarios for Consideration for IMT2020, NGMN Alliance, Frankfurt, Germany, 2016. [Online]. Available: https://www.ngmn.org

[31] L. Lu, X. Han, J. Li, J. Hua, and M. OuYang, “A review on the key issues for lithium-ion battery management in electric vehicles,” J. Power Sources, vol. 226, pp. 272–288, Mar. 2013.

[32] N. Piovesan and P. Dini, “Optimal direct load control of renewable powered small cells: A shortest path approach,” Internet Technol. Lett., vol. 1, no. 1, p. e7, 2017.

[33] N. Piovesan, M. Miozzo, and P. Dini, “Optimal direct load control of renewable powered small cells: Performance evaluation and bounds,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Apr. 2018, pp. 1–6.

[34] L. Bu¸soniu, R. Babuška, and B. De Schutter, Multi-Agent Reinforcement Learning: An Overview. Berlin, Germany: Springer, 2010, pp. 183–221.

[35] N. Morozs, T. Clarke, and D. Grace, “Distributed heuristically accelerated Q-learning for robust cognitive spectrum management in LTE cellular systems,” IEEE Trans. Mobile Comput., vol. 15, no. 4, pp. 817–825, Apr. 2016.

[36] X. Li et al., “Innovations through 5G-crosshaul applications,” in Proc. Eur. Conf. Netw. Commun. (EuCNC), Jun. 2016, pp. 382–387.

[37] EU EARTH: Energy Aware Radio and Network Technologies. (2010). D2.3: Energy Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown. [Online]. Available: www.iict-earth.eu

[38] M. Mezzavilla, M. Miozzo, M. Rossi, N. Baldo, and M. Zorzi, “A lightweight and accurate link abstraction model for system-level simulation of LTE networks in NS-3,” in Proc. ACM MSWiM, Paphos, Cyprus, Oct. 2012, pp. 55–60.

[39] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.

[40] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). New York, NY, USA: Springer-Verlag, 2006.

Marco Miozzo received the M.Sc. degree in telecommunication engineering from the University of Ferrara, Italy, in 2005, and the Ph.D. degree from the Technical University of Catalonia (UPC) in 2018. He was a Research Engineer of wireless networking with the Consorzio Ferrara Ricerche, Italy, and with the Department of Information Engineering, University of Padova, Italy. In June 2008, he joined the Centre Tecnològic de Telecomunicacions de Catalunya, where he has been involved in the development of an LTE module for the network simulator 3 in the framework of the LENA Project. He is currently collaborating with the EU founded H2020 SCAVENGE (MSCA ETN) as a Work Package 4 Leader and Student Supervisor. He participated in several research and development projects and he has been involved in 5G-Crosshaul, Flex5Gware, and SANSA Projects, working on environmental sustainable mobile networks with energy harvesting capabilities through learning techniques. His main research interests include sustainable mobile networks, green wireless networking, energy harvesting, and machine learning.

Nicola Piovesan received the B.Sc. degree in information engineering and the M.Sc. degree in telecommunication engineering from the University of Padova, Italy, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Department of Telematics Engineering, Technical University of Catalonia (UPC). In September 2016, he joined the Centre Tecnològic de Telecomunicacions de Catalunya as a Research Assistant in the framework of the EU H2020 MSCA Sustainable Cellular Networks Harvesting Ambient Energy Project. His current research interests include energy harvesting in wireless communications, machine learning, energy transfer systems, and resource allocation in wireless communication systems.

Paolo Dini received the Laurea degree in electronic engineering and the Ph.D. degree in information and communication science and technology from the Università di Roma “La Sapienza” in 2001 and 2005, respectively. He is a Senior Researcher with the Centre Tecnològic de Telecomunicacions de Catalunya. From 2001 to 2005, he was a Research Assistant with the Computer Science and Systems (DIS) Department and Information and Communication (INFOCOM) Department, Università di Roma “La Sapienza.” In 2005, he was with the Research Centre on Software Technologies as a Research Assistant. In 2005, he was also an Assistant Professor with the INFOCOM Department Lecturing the course of “Comunicazioni Elettriche.” He has an active participation in research projects funded by EU, national and regional government, as well as by private companies. He took part in over 20 research and development projects so far as a Principal Investigator, a WorkPackage Leader, and a Project Manager, or a Researcher. He is currently the Coordinator of the MSCA SCAVENGE European Training Network. His research interests encompass system modeling and optimization for machine learning, data analytics, and optimal control with particular emphasis on sustainability and energy saving, including energy harvesting networks and smart power grids. He was a recipient of the Silicon Valley Foundation Award in 2008 and 2011 for his research in heterogeneous mobile networks. He has also been serving as a TPC member in many international conferences and workshops and as a reviewer in several scientific journals of IEEE, Springer, Wiley, and Elsevier.