On-Demand AoI Minimization in Resource-Constrained Cache-Enabled IoT Networks With Energy Harvesting Sensors

Mohammad Hatami, *Graduate Student Member, IEEE*, Markus Leinonen, *Member, IEEE*, Zheng Chen, *Member, IEEE*, Nikolaos Pappas, *Senior Member, IEEE*, and Marian Codreanu, *Member, IEEE*

***Abstract*— We consider a resource-constrained IoT network, where multiple users make on-demand requests to a cache-enabled edge node to send status updates about various random processes, each monitored by an energy harvesting sensor. The edge node serves users' requests by deciding whether to command the corresponding sensor to send a fresh status update or retrieve the most recently received measurement from the cache. Our objective is to find the best actions of the edge node to minimize the average age of information (AoI) of the received measurements upon request, i.e., *average on-demand AoI*, subject to *per-slot transmission and energy constraints*. First, we derive a Markov decision process model and propose an iterative algorithm that obtains an optimal policy. Then, we develop an *asymptotically optimal low-complexity* algorithm – termed *relax-then-truncate* – and prove that it is optimal as the number of sensors goes to infinity. Simulation results illustrate that the proposed relax-then-truncate approach significantly reduces the average on-demand AoI compared to a request-aware greedy policy and a weighted AoI policy, and also depict that it performs close to the optimal solution even for moderate numbers of sensors.**

***Index Terms*— Age of information (AoI), energy harvesting (EH), constrained Markov decision process (CMDP).**

I. Introduction

INTERNET of Things (IoT) is a key technology in providing ubiquitous, intelligent networking solutions to create a smart society. In IoT sensing networks, sensors measure physical quantities (e.g., speed or pressure) and send the measurements to a destination for further processing. IoT networks are subject to stringent energy limitations, due to battery-powered sensors. This energy scarcity is often counteracted by the *energy harvesting* (EH) technology, relying on, e.g., solar or RF ambient sources. Moreover, reliable control actions in emerging time-critical IoT applications (e.g., drone control and industrial monitoring) require high *freshness* of information received by the destination. Such destination-centric information freshness can be quantified by the *age of information* (AoI) [2], [3]. These call for designing effective *AoI-aware status updating* procedures for IoT networks to provide the end users with timely status of remotely observed processes while accounting for the limited energy resources of EH sensors.

We consider a resource-constrained IoT sensing network consisting of multiple EH sensors, a cache-enabled edge node, and multiple users. Users are interested in timely status information about the random processes associated with physical quantities (e.g., speed or temperature), each measured by a sensor. We consider *request-based* status updating where the users demand for the status of physical quantities from the edge node which acts as a gateway. between the users and the sensors. The edge node is equipped with a *cache* that stores the most recently received *status update packet* from each sensor. Upon receiving request(s) for the status of a physical quantity, the edge node has two options to serve the requesting user(s): either command the corresponding sensor to send a fresh status update or use the aged measurement from the cache. The former enables serving a user with fresh measurement, yet consuming energy from the sensor's battery. The latter prevents the activation of the sensors for every request so that the sensors can utilize the sleep mode to save a considerable amount of energy [4], but the data received by the users becomes stale. Due to this intrinsic AoI-energy trade-off, the edge node must decide, in a farsighted fashion, when to provide the users with fresh status updates at the cost of sensors' energy expenditure and when to resort to use the cached (stale) measurements to save the sensors' batteries for the future requests.

In particular, the considered status updating network is subject to the following *energy and transmission constraints*. First, since the sensors rely only on the energy harvested from the environment, the sensors' batteries may be empty and thus they cannot send an update for each request. This *energy*

Manuscript received 26 January 2022; revised 24 July 2022; accepted 12 September 2022. Date of publication 22 September 2022; date of current version 18 November 2022. The work has been financially supported in part by Infotech Oulu, the Academy of Finland (grant 323698), and Academy of Finland 6G Flagship Program (grant 346208). Mohammad Hatami would like to acknowledge the support of Nokia Foundation. The work of Markus Leinonen has also been financially supported in part by the Academy of Finland (grant 340171 and 319485). The work of Zheng Chen and Nikolaos Pappas have been supported in part by the Swedish Research Council (VR), ELLIIT, and CENIIT. An earlier version of this paper was presented in part at the IEEE International Symposium on Information Theory, Espoo, Finland, in 2022 [DOI: 10.1109/ISIT50566.2022.9834773]. The associate editor coordinating the review of this article and approving it for publication was C. Li. *(Corresponding author: Mohammad Hatami.)*

Mohammad Hatami and Markus Leinonen are with the Centre for Wireless Communications–Radio Technologies, University of Oulu, 90014 Oulu, Finland (e-mail: mohammad.hatami@oulu.fi; markus.leinonen@oulu.fi).

Zheng Chen is with the Department of Electrical Engineering, Linköping University, 58183 Linköping, Sweden (e-mail: zheng.chen@liu.se).

Nikolaos Pappas is with Department of Computer and Information Science, Linköping University, 58183 Linköping, Sweden (e-mail: nikolaos.pappas@liu.se).

Marian Codreanu is with Department of Science and Technology, Linköping University, 58183 Linköping, Sweden (e-mail: marian.codreanu@liu.se).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2022.3208873.

Digital Object Identifier 10.1109/TCOMM.2022.3208873

causality induces an inherent *per-slot energy constraint*. Second, motivated by the limited amount of radio resources (e.g., bandwidth, time-frequency resource blocks), only a limited number of sensors can send fresh status updates to the edge node at each time slot, imposing a *per-slot transmission constraint*.

The objective of our network design is to keep the freshness of information at the users as small as possible, subject to the constraints in the system. To this end, we use the concept of *on-demand AoI* [5] that quantifies the freshness of information at the users restricted to the users' request instants. We aim to find an *optimal policy*, i.e., the best action of the edge node at each time slot that minimizes the average on-demand AoI over all the sensors and the users subject to the per-slot transmission and energy constraints.

We first cast the problem as an average-cost Markov decision process (MDP) for which the RVIA [6, Section 8.5.5] is used to obtain an optimal policy. Then, since the complexity of finding an optimal policy increases exponentially in the number of sensors, we propose an asymptotically optimal low-complexity algorithm – termed *relax-then-truncate* – and show that it performs close to the optimal solution.

A. Contributions

The main contributions of our paper are as follows:

- We consider on-demand AoI minimization problem in a multi-user multi-sensor IoT EH network subject to per-slot transmission and energy constraints. The problem is formulated as an average-cost MDP for which the RVIA is used to obtain an optimum policy.
- To deal with massive IoT scenarios, we propose a sub-optimal low-complexity algorithm whose complexity increases linearly in the number of sensors. In particular, we relax the per-slot transmission constraint into a time average constraint, model the relaxed problem as a constrained MDP (CMDP), obtain an optimal relaxed policy, and propose an online truncation procedure to ensure that the transmission constraint is satisfied at each time slot.
- We analytically find an upper bound for the difference between the average cost obtained by the proposed relax-then-truncate approach and the average cost obtained by an optimal policy. Then, we show that the relax-then-truncate approach is asymptotically optimal as the number of sensors goes to infinity.
- Numerical experiments are conducted to analyze the performance of the proposed relax-then-truncate approach and show that it significantly reduces the average on-demand AoI as compared to a request-aware greedy policy and a weighted AoI policy. Interestingly, the proposed algorithm performs close to the optimal solution for moderate numbers of sensors.

Our model is representative in IoT networks and highly relevant to resource-constrained IoT scenarios with a massive number of devices, a setup of paramount importance in practice, because the number of IoT sensors can grow to large numbers in emerging IoT applications. *To the best of our knowledge, this work is the first one that proposes an asymptotically optimal low-complexity algorithm for minimizing on-demand AoI in an IoT network with multiple EH sensors*.

B. Related Work

AoI-optimal scheduling has attracted considerable research interest over the last few years [5], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38]. Particularly, a popular approach is to model the problem as an MDP and find an optimal policy by using model-based reinforcement learning (RL) methods [5], [7], [11], [12], [13], [14], [22], [23], [25], [26], [27], [29], [31], [37], [38], e.g., relative value iteration algorithm (RVIA), and/or model-free RL methods [5], [11], [12], [15], [16], [20], [28], [29], [30], [31], [33], e.g., (deep) Q-learning.

In [7], the authors proposed AoI-optimal scheduling algorithms for a broadcast network where a base station is updating the users on random information arrivals under a transmission capacity constraint. In [8], the authors developed low-complexity scheduling algorithms, including a Whittle's index policy, and derived performance guarantees for a broadcast network. In [9] and [10], the optimality of the Whittle's index policy has been investigated for the AoI minimization problem where a central entity schedules a number of users among the total available users for transmission over unreliable channels. In [11], the authors studied AoI-optimal scheduling under a constraint on the average number of transmissions where the source sends status updates to a destination (user) over an error-prone channel. The authors in [12] extended [11] to a multi-user setting, where the source has to decide not only when to transmit but also to which user. In [13], the authors proposed an asymptotically optimal algorithm for the AoI-optimal scheduling problem under both bandwidth and average power constraints in a wireless network with time-varying channel states. In [14], the authors studied AoI minimization problem in a multi-source relaying system under per-slot transmission and average resource constraints. In [15], the authors used a deep RL framework to minimize the weighted average AoI plus energy cost for online cache updating in IoT networks under dynamic content popularity. In [16], the authors developed a multi-agent RL framework for cache updating in IoT sensing networks where the objective is to minimize the weighted average AoI plus energy cost and fronthaul traffic loads.

Different from [7], [8], [9], [10], [11], [12], [13], [14], another line of research [17], [20], [22], [23], [24], [25], [26], [27], [30] focused on the class of problems where the sources are powered by *energy harvested from the environment*, i.e., investigating AoI-optimal scheduling policies subject to the energy causality constraint at the source(s). The works [17], [20], [22], [23], [24], [25], [26], [27] studied AoI-optimal scheduling in single-sensor EH networks where the sensor sends time-sensitive information to the user(s). In [17], the authors obtained an AoI-optimal policy for the sensor's sampling instants by assuming known EH statistics. In [18], the authors derived age-optimal online policies for an EH sensor having a unit-sized battery or infinite battery using renewal theory. In [19], the authors derived age-optimal policies for an EH sensor with a finite-sized battery. The authors of [20] studied AoI-optimal policies under an erasure channel with retransmissions in a system where the EH and channel statistics are either unknown or known. In [21], the authors studied online status updating under updating erasures for the cases where no feedback or perfect feedback is available to the source. In [22], AoI-optimal scheduling was studied in a system where the sensor uses multiple transmission modes. The work [23] investigated age-optimal scheduling for a cognitive radio EH system. The authors of [24] studied AoI-optimal scheduling under stability constraints in a multiple access channel with two heterogeneous nodes (including an EH node) transmitting to a common destination. In [25], the sensor monitors a stochastic process and tracks its evolution and thereby, a modified definition of AoI is proposed to account for the discrepancy in the remote destination. In [26], the monitoring node (sensor) collects status updates from multiple heterogeneous information sources. In [27], the authors studied AoI-optimal scheduling for a wireless powered communication system under the costs of generating status updates at the sensor nodes. In [28], the authors utilized the Q-learning algorithm to obtain an age-optimal policy for a scenario where the EH source samples and forwards the measurements to a monitoring center over a millimeter-wave channel. In [29], the authors studied age-optimal status updating over a time-varying wireless link. In [30], the authors developed a deep RL algorithm for minimizing the average age of correlated information in an IoT network with multiple correlated EH sensors. In [31], deep RL was used to minimize AoI in a multi-node monitoring system, in which the sensors are powered through wireless energy transfer by the destination. In [32], the authors proposed several harvesting-aware energy management policies for solar-powered IoT devices that asynchronously send status updates to a gateway device.

The majority of the literature that considers AoI minimization, including all the above ones, assume that the time-sensitive information of the source(s) is needed at the destination *at all time moments*. However, in many applications, a user demands for fresh status updates only when it needs such timely information. To account for such information freshness driven by users' requests, we introduced the concept of on-demand AoI in [5], [33]. In these works and a follow-up work [34], the main focus was on-demand AoI minimization in an IoT network with multiple *decoupled* EH sensors. On the contrary, the main distinctive feature of this paper is to study optimal scheduling under per-slot *transmission constraints*. Only a few works have investigated a concept similar to the on-demand AoI, yet in different frameworks. The work [35] introduced effective AoI under a generic request-response model where a server provides time-sensitive data to the users. An information update system with a user that pulls information from servers was investigated in [36]. However, contrary to our paper, [35], [36] do not consider energy limitations at the source nodes and the frameworks are fundamentally different. In [37] and [38], the authors introduced the AoI at query (QAoI) and developed an MDP-based policy iteration method to find an optimal policy that minimizes the average QAoI considering an energy-constrained sensor that is queried to send updates to an edge node under limited transmission opportunities. The main difference between our paper and [37], [38] is that we consider IoT networks with multiple EH sensors. Moreover, the on-demand AoI metric is not the same as QAoI in [37] and [38].

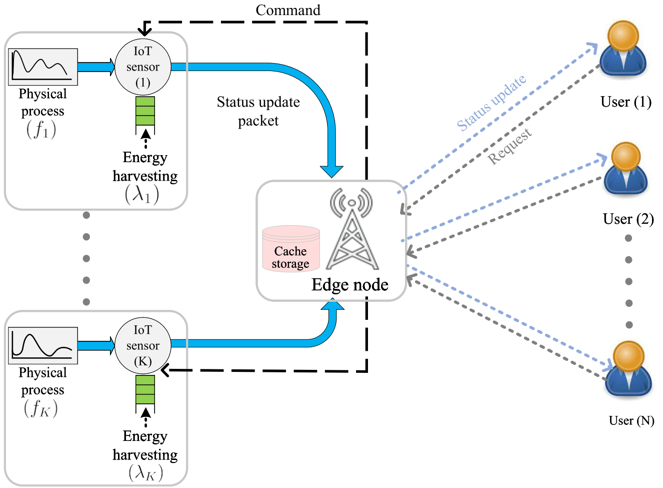


Fig. 1. A multi-user multi-sensor IoT sensing network consisting of K EH sensors, an edge node, and N users. The end users are interested in timely status update information of the physical processes measured by the sensors.

II. System Model and Problem Formulation

A. Network Model

We consider a multi-user multi-sensor IoT sensing network that consists of a set $\mathcal{K} = \{1, \ldots, K\}$ of K energy harvesting (EH) sensors, an edge node (a gateway), and a set $\mathcal{N} = \{1, \ldots, N\}$ of N users, as depicted in Fig. 1. Users are interested in timely status information about random processes associated with physical quantities f_k, e.g., speed or temperature, each of which is independently measured by sensor $k \in \mathcal{K}$. We assume that the sensors measure different physical quantities. We consider *request-based* status updating, where the users send requests *on demand* for obtaining status of quantities f_k, $k \in \mathcal{K}$. When a request for the physical quantity f_k is generated at the user side, the associated sensor k may send a *status update packet* that contains the measured value of the monitored process and a time stamp of the generated sample. We assume that there is no direct link between the users and the sensors, i.e., the users receive the status updates only via the edge node. The edge node provides an interface for the users to communicate with IoT sensors.[1]

We consider a time-slotted system with slots indexed by $t \in \mathbb{N}$. At the beginning of slot t, users send requests for the status of physical quantities f_k to the edge node. Let $r_{k,n}(t) \in \{0, 1\}$, $t = 1, 2, \ldots$, denote the random process

[1] Particularly, our system model can represent a scenario where the (low-power) energy harvesting sensors are located in a remote area for which the users are out of the sensors' communication range. Furthermore, considering a control node (e.g., gateway) between data generators (e.g., sensors) and devices requesting data (e.g., users) is common in IoTs (see e.g., [4], [15]).

of requesting the status of f_k by user n; $r_{k,n}(t) = 1$ if the status of f_k is requested by user $n \in \mathcal{N}$ at slot t and $r_{k,n}(t) = 0$ otherwise. The requests are independent across the users, sensors, and time slots. Let $p_{k,n}$ be the probability that the status of f_k is requested by user n at each slot, i.e., $\Pr\{r_{k,n}(t) = 1\} = p_{k,n}$. Note that a user might request for multiple physical quantities at each time slot. Moreover, there can be multiple users requesting for f_k at each slot; $r_k(t) = \sum_{n=1}^{N} r_{k,n}(t) \in \{0, 1, \ldots, N\}$ indicates the number of requests for f_k at slot t. We assume that all requests that arrive at the beginning of slot t are handled by the edge node during the same slot t. Note that while the communications between the edge node and the users are assumed to be error-free,[2] the transmissions from the sensors to the edge node are prone to errors, as detailed in Section II-C.

The edge node is equipped with a *cache* of size K that stores the most recently received status update packet from each sensor. Upon receiving a request for the status of f_k at slot t, the edge node has two options to serve the request: 1) command sensor k to send a fresh status update, or 2) use the previous measurement from the cache. We denote the *command action of the edge node* at slot t by $a_k(t) \in \{0, 1\}$; $a_k(t) = 1$ if the edge node commands sensor k to send an update and $a_k(t) = 0$ otherwise.

We consider that, due to limited amount of radio resources (e.g., time-frequency resource blocks), no more than $M \le K$ sensors can transmit status updates to the edge node within each slot. This *transmission constraint* imposes a limitation to the number of commands as

$$\sum_{k=1}^{K} a_k(t) \le M, \forall t. \tag{1}$$

We refer to M as the *transmission budget* hereinafter.

B. Energy Harvesting Sensors

We assume that the sensors harvest energy from the environment for sustainable operation. We model the energy arrivals at the sensors as independent Bernoulli processes[3] with intensities λ_k, $k \in \mathcal{K}$. This characterizes the discrete nature of the energy arrivals in a slotted-time system, i.e., at each slot, a sensor either harvests one unit of energy or not (see, e.g., [17], [25], [40]). We denote the *energy arrival process* of sensor k by $e_k(t) \in \{0, 1\}$, $t = 1, 2, \ldots$. Therefore, during each time slot, sensor k harvests one unit of energy with probability λ_k,[4] i.e., $\Pr\{e_k(t) = 1\} = \lambda_k$, $\forall t$. For sensor k, the harvested energy is stored in a battery with a finite capacity B_k. We denote the battery level of sensor k at the beginning of slot t by $b_k(t)$, where $b_k(t) \in \{0, \ldots, B_k\}$.

We assume that measuring and transmitting a status update from each sensor to the edge node consumes one unit of energy, i.e., the energy unit is normalized so that each status update requires one unit of energy (see, e.g., [5], [18], [21], [25], [30], [40], [41]). Once sensor k receives a command from the edge node (i.e., $a_k(t) = 1$), the sensor sends a status update if its battery is non-empty (i.e., $b_k(t) \ge 1$). We denote the *action of sensor* k at slot t by $d_k(t) \in \{0, 1\}$; $d_k(t) = 1$ if sensor k sends a status update to the edge node and $d_k(t) = 0$ otherwise. Hence, the sensor's action, the edge node's action, and the battery level of the sensor are interrelated as

$$d_k(t) = a_k(t)\mathbb{1}_{\{b_k(t) \ge 1\}}, \tag{2}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Note that $d_k(t)$ in (2) determines the energy expenditure of sensor k at slot t. It is also worth noting that by (2), we have $d_k(t) \le a_k(t)$, and consequently, (1) implies that $\sum_{k=1}^{K} d_k(t) \le M$ for all slots; hence, the name transmission constraint for (1).

Finally, using $b_k(t)$, $d_k(t)$, and $e_k(t)$, the evolution of the battery level of sensor k is given by

$$b_k(t+1) = \min\{b_k(t) + e_k(t) - d_k(t), B_k\}. \tag{3}$$

C. Communication Between the Edge Node and the Sensors

We consider an *error-free* binary/single-bit *command* link from the edge node to each sensor (see e.g., [42], [43]), and an *error-prone* wireless communication link from each sensor to the edge node. If a sensor sends a status update packet to the edge node, the transmission through the wireless link can be either *successful* or *failed*. Let $w_k(t) = 1$ denote the event that a status update from sensor k has been successfully received by the edge node at slot t. Otherwise, $w_k(t) = 0$ which accounts for both the cases that either 1) sensor k sends a status update but the transmission is failed, or 2) the sensor does not send a status update. Let ξ_k be the conditional probability that given that sensor k transmits a status update, it is successfully received by the edge node, i.e., $\Pr\{w_k(t) = 1 \mid d_k(t) = 1\} = \xi_k$, $k \in \mathcal{K}$, $t = 1, 2, \ldots$. Thus, ξ_k represents the *transmit success probability* of the link from sensor k to the edge node.

D. On-Demand Age of Information

To measure the freshness of information seen by the users in our request-based status updating system, we use the notion of age of information (AoI) [2] and define *on-demand AoI* [5]. In contrast to AoI that measures the freshness of information at every slot, on-demand AoI quantifies *the freshness of information at the users' request instants (only)*.

Let $\Delta_k(t)$ be the AoI about the physical quantity f_k at the edge node at the beginning of slot t, i.e., the number of slots elapsed since the generation of the most recently received status update packet from sensor k. Let $u_k(t)$ denote the most recent slot in which the edge node received a status update packet from sensor k, i.e.,

[2] This assumption is invoked by the fact that the edge node has access to power (e.g., a base station connected to a fixed power grid), whereas the sensors rely only on the energy harvested from the environment. Moreover, in some applications, the edge node may be connected to the users via a wired or fiber connection or via internet.

[3] The i.i.d. energy arrivals are commonly considered in the literature; see e.g., [4], [25], [26], [30]. Alternatively, one can consider temporally correlated energy arrivals, such as discrete-Markovian modeling [39], which are beyond the scope of this work.

[4] The scenario where the sensors access to adequate power (e.g., connected to a fixed power grid) so that they can send a status update at every slot is a special case of our system model, i.e., a unit energy arrival at every slot, $\lambda_k = 1$.

$u_k(t) = \max\{t' | t' < t, w_k(t') = 1\}$. Thus, the AoI about f_k is given by the random process $\Delta_k(t) \triangleq t - u_k(t)$.

We make a common assumption (see e.g., [5], [9], [11], [12], [14], [20], [22], [23], [25], [26], [27], [30], [33]) that $\Delta_k(t)$ is upper-bounded by a finite value $\Delta^{\max}$, i.e., $\Delta_k(t) \in \{1, 2, \ldots, \Delta^{\max}\}$. Besides tractability, this accounts for the fact that once the available measurement about f_k becomes excessively stale, further counting would be irrelevant. At each slot, the AoI about f_k drops to one if the edge node receives a status update from sensor k; otherwise, it increases by one. Therefore, $\Delta_k(t)$ evolves as

$$\Delta_k(t+1) = \begin{cases} 1, & \text{if } w_k(t) = 1, \\ \min\{\Delta_k(t) + 1, \Delta^{\max}\}, & \text{if } w_k(t) = 0. \end{cases} \tag{4}$$

The compact form of (4) is written as $\Delta_k(t+1) = \min\{(1 - w_k(t))\Delta_k(t) + 1, \Delta^{\max}\}$.

We define on-demand AoI for a sensor-user pair (k, n) at slot t as the sampled version of (4) where the sampling is controlled by the request process $r_{k,n}(t)$, i.e.,

$$\begin{aligned} \Delta^{\mathrm{OD}}_{k,n}(t) &\triangleq r_{k,n}(t)\Delta_k(t+1) \\ &= r_{k,n}(t) \min\{(1 - d_k(t))\Delta_k(t) + 1, \Delta^{\max}\}. \end{aligned} \tag{5}$$

In (5), since the requests come at the beginning of slot t and the edge node sends measurements to the users at the end of the same slot, $\Delta_k(t+1)$ is the AoI about f_k seen by the users.

E. State Space, Action Space, Policy, and Cost Function

1) State: Let $s_k(t) \in \mathcal{S}_k$ denote the state associated with sensor k at slot t, which is defined as $s_k(t) = (r_k(t), b_k(t), \Delta_k(t))$, where $r_k(t) \in \{0, 1, \ldots, N\}$ indicates the number of requests for f_k, $b_k(t) \in \{0, 1, \ldots, B_k\}$ is the battery level,[5] and $\Delta_k(t) \in \{1, 2, \ldots, \Delta^{\max}\}$ is the AoI about f_k at the edge node; $\mathcal{S}_k$ is the per-sensor state space with dimension $|\mathcal{S}_k| = (N + 1)(B_k + 1)\Delta^{\max}$. The state of the system at slot t is expressed as $\mathbf{s}(t) = (s_1(t), \ldots, s_K(t)) \in \mathcal{S}$, $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_K$; the state space $\mathcal{S}$ has a finite dimension $|\mathcal{S}| = \prod_{k=1}^{K}(N + 1)(B_k + 1)\Delta^{\max}$.

2) Action: As discussed in Section II-A, the edge node decides at each slot whether to command sensor k to send a fresh status update (and update the cache) or not, i.e., $a_k(t) \in \mathcal{A}_k = \{0, 1\}$, where $\mathcal{A}_k$ is the per-sensor action space. The action of the edge node at slot t is given by a K-tuple $\mathbf{a}(t) = \big(a_1(t), \ldots, a_K(t)\big) \in \mathcal{A}$ with action space $\mathcal{A} = \big\{(a_1, \ldots, a_K) \mid a_k \in \mathcal{A}_k, \sum_{k=1}^{K} a_k \leq M\big\}$. The action space dimension is $|\mathcal{A}| = \sum_{m=0}^{M} \binom{K}{m}$. It is worth stressing that the action space $\mathcal{A}$ considers the transmission constraint (1) in its definition. Additionally, we define the *relaxed* action space that does not consider the transmission constraint (1) as $\mathcal{A}_{\mathrm{R}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K = \{0, 1\}^K$, which has the dimension $|\mathcal{A}_{\mathrm{R}}| = 2^K$.

3) Policy: A policy π determines an action at a given state. A randomized policy is a mapping from state $\mathbf{s} \in \mathcal{S}$ to a *probability distribution* $\pi(\mathbf{a}|\mathbf{s}) : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, $\sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a}|\mathbf{s}) = 1$, of choosing each possible action $\mathbf{a} \in \mathcal{A}$. A deterministic policy is a special case where, in each state $\mathbf{s}$, $\pi(\mathbf{a}|\mathbf{s}) = 1$ for some $\mathbf{a}$; with a slight abuse of notation, we use $\pi(\mathbf{s})$ to denote the action taken in state $\mathbf{s}$ by a deterministic policy π. In addition, we define a (relaxed) policy as $\pi_{\mathrm{R}} : \mathcal{S} \times \mathcal{A}_{\mathrm{R}} \rightarrow [0, 1]$ and a per-sensor policy as $\pi_k : \mathcal{S}_k \times \mathcal{A}_k \rightarrow [0, 1]$.

4) Cost Function: We consider a cost function that incurs a penalty with respect to the staleness of a status update requested and received by a user. Accordingly, we define the cost associated with user n and sensor k at slot t as the on-demand AoI for the sensor-user pair (k, n), i.e., $\Delta^{\mathrm{OD}}_{k,n}(t)$ defined in (5). Then, the per-sensor cost at slot t is expressed as

$$\begin{aligned} c_k(t) &= \sum_{n=1}^{N} \Delta^{\mathrm{OD}}_{k,n}(t) = \sum_{n=1}^{N} r_{k,n}(t)\Delta_k(t+1) \\ &= r_k(t)\Delta_k(t+1). \end{aligned} \tag{6}$$

Remark 1: Note that, due to the multiplicative factor $r_k(t)$, (6) accounts for the number of requests for each physical quantity at each slot, i.e., the more the requests for f_k, the more important the corresponding freshness becomes. Particularly, when the status of f_k is not requested by any user at slot t, i.e., $r_k(t) = 0$, the immediate cost becomes $c_k(t) = 0$.

F. Problem Formulation

For the considered system, the energy and transmission constraints pose limitations on when and how often a new status update can be generated at each sensor, which in turn affect the on-demand AoI. Our objective is to keep the on-demand AoI as small as possible, subject to the constraints in the system. Formally, for a given policy π, we define the average cost as *the average on-demand AoI over all sensors and users,* i.e.,

$$\bar{C}_\pi \triangleq \lim_{T \to \infty} \frac{1}{NKT} \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{E}_\pi[c_k(t) \mid \mathbf{s}(0)], \tag{7}$$

where $\mathbb{E}_\pi[\cdot]$ is the (conditional) expectation when the policy π is applied to the system and $\mathbf{s}(0) = \big(s_1(0), \ldots, s_K(0)\big)$ is the initial state.[6] We aim to find an optimal policy $\pi^\star$ that achieves the minimum average cost, i.e.,

$$(\mathbf{P1}) \quad \pi^\star \in \arg\min_{\pi} \bar{C}_\pi. \tag{8}$$

III. MDP Modeling and Optimal Policy

In this section, we model the problem (**P1**) as an MDP and propose a value iteration algorithm that finds an optimal policy $\pi^\star$.

[5] We assume that there are low-cost ("1-bit") control channels between the sensors and the edge node so that a sensor updates the edge node whenever its battery level changes (increases) due to the harvested energy.

[6] As shown in Proposition 2, the minimum average cost is independent of the initial state, thus, we omit the initial state henceforth.

A. MDP Modeling

The MDP is defined by the tuple $\big(\mathcal{S}, \mathcal{A}, \Pr(\mathbf{s}(t+1)|\mathbf{s}(t), \mathbf{a}(t)), c(\mathbf{s}(t), \mathbf{a}(t))\big)$. The state space $\mathcal{S}$ and the action space $\mathcal{A}$ were defined in Section II-E. The cost function $c(\mathbf{s}(t), \mathbf{a}(t))$ represents the cost of taking action $\mathbf{a}(t)$ in state $\mathbf{s}(t)$, which is given by $c(\mathbf{s}(t), \mathbf{a}(t)) = \frac{1}{NK}\sum_{k=1}^{K} c_k(s_k(t), a_k(t))$, where the per-sensor cost $c_k(s_k(t), a_k(t))$ is calculated using (6), i.e.,

$$\begin{aligned} &c_k(s_k(t), a_k(t)) \\ &= r_k(t)\big[\xi_k \min\big\{\big(1 - a_k(t)\mathbb{1}_{\{b_k(t) \ge 1\}}\big) \\ &\quad \Delta_k(t) + 1, \Delta^{\max}\big\} + (1-\xi_k)\min\left\{\Delta_k(t)+1, \Delta^{\max}\right\}\big]. \end{aligned} \tag{9}$$

The state transition probability $\Pr(\mathbf{s}(t+1)|\mathbf{s}(t), \mathbf{a}(t))$ maps a state-action pair at slot t onto a distribution of states at slot $t+1$. The probability of transition from current state $\mathbf{s}(t) = (s_1(t), \ldots, s_K(t))$ to next state $\mathbf{s}(t+1) = (s_1(t+1), \ldots, s_K(t+1))$ under action $\mathbf{a}(t) = (a_1(t), \ldots, a_K(t))$ factorizes as

$$\Pr\left(\mathbf{s}(t+1) \mid \mathbf{s}(t), \mathbf{a}(t)\right) \overset{(a)}{=} \prod_{k=1}^{K} \Pr\left(s_k(t+1) \mid s_k(t), a_k(t)\right),$$

where (a) follows from the fact that given action $\mathbf{a}$, the state associated with each sensor (i.e., the per-sensor state) evolves independently from the other sensors. Above, the per-sensor state transition probability $\Pr\left(s_k(t+1) \mid s_k(t), a_k(t)\right)$ gives the probability of transition from per-sensor state $s_k(t) = (r_k, b_k, \Delta_k)$ to next per-sensor state $s_k(t+1) = (r'_k, b'_k, \Delta'_k)$ under action $a_k(t) = a_k$, and it is expressed as

$$\begin{aligned} &\Pr\left(s_k(t+1) \mid s_k(t), a_k(t)\right) \\ &\quad \triangleq \Pr\left(r'_k, b'_k, \Delta'_k \mid r_k, b_k, \Delta_k, a_k\right) \\ &\quad \overset{(a)}{=} \underbrace{\Pr\left(r'_k \mid r_k, b_k, \Delta_k, a_k\right)}_{\overset{(b)}{=}\Pr(r'_k)} \underbrace{\Pr\left(b'_k \mid r_k, b_k, \Delta_k, a_k, r'_k\right)}_{\overset{(c)}{=}\Pr\left(b'_k|b_k, a_k\right)} \\ &\quad \times \underbrace{\Pr\left(\Delta'_k \mid r_k, b_k, \Delta_k, a_k, r'_k, b'_k\right)}_{\overset{(d)}{=}\Pr\left(\Delta'_k|b_k, \Delta_k, a_k\right)} \\ &\quad = \Pr(r'_k)\Pr\left(b'_k \mid b_k, a_k\right)\Pr\left(\Delta'_k \mid b_k, \Delta_k, a_k\right), \end{aligned} \tag{10}$$

where (a) follows from the chain rule, (b) follows from the independence between the request process and the other random variables, (c) follows because, given current battery level b_k and action a_k, next battery level b'_k is independent of the requests and the current AoI, and (d) follows since given b_k, Δ_k, and a_k, the next value of AoI Δ'_k can be obtained (see (4)). The probabilities in (10) are calculated in the following.

The random variable $r'_k = \sum_n r'_{k,n}$ is a sum of independent Bernoulli trials that are not necessarily identically distributed. Therefore, it has a Poisson binomial distribution [44] as

$$\Pr(r'_k) = \begin{cases} \prod_{n=1}^{N}(1-p_{k,n}), & r'_k = 0, \\ \sum_{n=1}^{N} p_{k,n}\prod_{m \neq n}(1-p_{k,m}), & r'_k = 1, \\ \cdots & \cdots \\ \prod_{n=1}^{N} p_{k,n}, & r'_k = N. \end{cases} \tag{11}$$

At each slot, sensor k consumes one unit of energy for sending a status update (i.e., when $a_k(t) = 1$ and $b_k(t) \ge 1$) and harvests one unit of energy with probability λ_k, thus, we have

$$\Pr(b'_k \mid b_k < B_k, a_k = 0) = \begin{cases} \lambda_k, & b'_k = b_k + 1, \\ 1 - \lambda_k, & b'_k = b_k, \\ 0, & \text{otherwise}. \end{cases} \tag{12a}$$

$$\Pr(b'_k \mid b_k = 0, a_k = 1) = \begin{cases} \lambda_k, & b'_k = 1, \\ 1 - \lambda_k, & b'_k = 0, \\ 0, & \text{otherwise}. \end{cases} \tag{12b}$$

$$\Pr(b'_k \mid b_k = B_k, a_k = 0) = \mathbb{1}_{\{b'_k = B_k\}}, \tag{12c}$$

$$\Pr(b'_k \mid b_k \ge 1, a_k = 1) = \begin{cases} \lambda_k, & b'_k = b_k, \\ 1 - \lambda_k, & b'_k = b_k - 1, \\ 0, & \text{otherwise}. \end{cases} \tag{12d}$$

According to (4) and (2), given current battery level b_k, AoI Δ_k, and action a_k, the next value of AoI Δ'_k can be obtained. Thus, we have

$$\Pr(\Delta'_k \mid b_k, \Delta_k, a_k = 0) = \mathbb{1}_{\{\Delta'_k = \min\{\Delta_k + 1, \Delta^{\max}\}\}}, \tag{13a}$$

$$\begin{aligned} &\Pr(\Delta'_k \mid b_k \ge 1, \Delta_k, a_k = 1) \\ &= \begin{cases} \zeta_k, & \Delta'_k = 1, \\ 1 - \zeta_k, & \Delta'_k = \min\left\{\Delta_k(t) + 1, \Delta^{\max}\right\}, \\ 0, & \text{otherwise}. \end{cases} \end{aligned} \tag{13b}$$

$$\Pr(\Delta'_k \mid b_k = 0, \Delta_k, a_k = 1) = \mathbb{1}_{\{\Delta'_k = \min\{\Delta_k + 1, \Delta^{\max}\}\}}. \tag{13c}$$

B. Optimal Policy

We propose an iterative algorithm that obtains an optimal policy $\pi^\star$ for (**P1**). we first define the accessibility condition for an MDP and prove that our MDP modeling in Section III-A satisfies this condition. Then, we present a proposition that characterizes an optimal policy $\pi^\star$ for (**P1**).

Definition 1: An MDP is *weakly communicating* (or *weakly accessible*) if the set of states can be partitioned into two subsets $\mathcal{S}_{\text{t}}$ and $\mathcal{S}_{\text{c}}$ such that: (i) all states in $\mathcal{S}_{\text{t}}$ are transient under every stationary policy and (ii) every two states in $\mathcal{S}_{\text{c}}$ can be reached from each other under some stationary policy [45, Definition 4.2.2]. In particular, an MDP is communicating (or accessible) if every two states can be reached from each other under some stationary policy.

Proposition 1: The MDP defined in Section III-A is weakly communicating.

Proof: The proof is presented in Appendix VII-A. □

Algorithm 1 RVIA That Obtains Optimal Policy $\pi^\star$

```
Initialize $V(\mathbf{s})\leftarrow 0$, $h(\mathbf{s})\leftarrow 0$, $\forall \mathbf{s}\in\mathcal{S}$, choose a reference state $\mathbf{s}_{\mathrm{ref}}\in\mathcal{S}$ and a small $\theta>0$
repeat
    for $\mathbf{s}\in\mathcal{S}$ do
        $V_{\mathrm{tmp}}(\mathbf{s})\leftarrow\min_{\mathbf{a}\in\mathcal{A}}[c(\mathbf{s},\mathbf{a})+\sum_{\mathbf{s}'\in\mathcal{S}}\Pr(\mathbf{s}'\,|\,\mathbf{s},\mathbf{a})h(\mathbf{s}')]$
    end for
    $\delta\leftarrow\max_{\mathbf{s}\in\mathcal{S}}(V_{\mathrm{tmp}}(\mathbf{s})-V(\mathbf{s}))-\min_{\mathbf{s}\in\mathcal{S}}(V_{\mathrm{tmp}}(\mathbf{s})-V(\mathbf{s}))$
    $V(\mathbf{s})\leftarrow V_{\mathrm{tmp}}(\mathbf{s})$, for all $\mathbf{s}\in\mathcal{S}$
    $h(\mathbf{s})\leftarrow V(\mathbf{s})-V(\mathbf{s}_{\mathrm{ref}})$, for all $\mathbf{s}\in\mathcal{S}$.
until $\delta<\theta$
$\pi^\star(\mathbf{s})=\arg\min_{\mathbf{a}\in\mathcal{A}}[c(\mathbf{s},\mathbf{a})+\sum_{\mathbf{s}'\in\mathcal{S}}\Pr(\mathbf{s}'\mid\mathbf{s},\mathbf{a})h(\mathbf{s}')]$, for all $\mathbf{s}\in\mathcal{S}$
```

Proposition 2: The optimal average cost achieved by an optimal policy $\pi^\star$, denoted by $\bar{C}^\star$ (i.e., $\bar{C}^\star=\bar{C}_{\pi^\star}$), is independent of the initial state $\mathbf{s}(0)$ and satisfies the Bellman's equation, i.e., there exists $h(\mathbf{s})$, $\mathbf{s}\in\mathcal{S}$, such that

$$\bar{C}^\star+h(\mathbf{s})=\min_{\mathbf{a}\in\mathcal{A}}\left[c(\mathbf{s},\mathbf{a})+\sum_{\mathbf{s}'\in\mathcal{S}}\Pr(\mathbf{s}'|\mathbf{s},\mathbf{a})h(\mathbf{s}')\right],\quad \mathbf{s}\in\mathcal{S}. \tag{14}$$

Further, an optimal action taken in state $\mathbf{s}$ is given by

$$\pi^\star(\mathbf{s})\in\arg\min_{\mathbf{a}\in\mathcal{A}}\left[c(\mathbf{s},\mathbf{a})+\sum_{\mathbf{s}'\in\mathcal{S}}\Pr(\mathbf{s}'|\mathbf{s},\mathbf{a})h(\mathbf{s}')\right],\quad \mathbf{s}\in\mathcal{S}. \tag{15}$$

Proof: By Proposition 1, the weak accessibility condition holds, thus, by [45, Prop. 4.2.6], there exists an optimal stationary (possibly randomized) policy, and by [45, Prop. 4.2.3], the optimal average cost $\bar{C}^\star$ is independent of the initial state. Furthermore, by [45, Prop. 4.2.1], if we can find such $\bar{C}^\star$ and $h(\mathbf{s})$ that satisfy (15), then (16) expresses an optimal policy for the problem. □

An optimal policy $\pi^\star$ can be found by turning the Bellman's optimality equation (14) into an iterative procedure, called relative value iteration algorithm (RVIA) [6, Section 8.5.5]. Particularly, at each iteration $i=0,1,\ldots$, we have

$$V^{(i+1)}(\mathbf{s})=\min_{\mathbf{a}\in\mathcal{A}}\left[c(\mathbf{s},\mathbf{a})+\sum_{\mathbf{s}'\in\mathcal{S}}\Pr(\mathbf{s}'\mid\mathbf{s},\mathbf{a})h^{(i)}(\mathbf{s}')\right], \tag{16}$$

$$h^{(i+1)}(\mathbf{s})=V^{(i+1)}(\mathbf{s})-V^{(i+1)}(\mathbf{s}_{\mathrm{ref}}), \tag{17}$$

where $\mathbf{s}_{\mathrm{ref}}\in\mathcal{S}$ is an arbitrary reference state. For any initialization $V^{(0)}(\mathbf{s})$, the sequences $\{h^{(i)}(\mathbf{s})\}_{i=1,2,\ldots}$ and $\{V^{(i)}(\mathbf{s})\}_{i=1,2,\ldots}$ converge [6, Section 8.5.5], i.e., $\lim_{i\to\infty}h^{(i)}(\mathbf{s})=h(\mathbf{s})$ and $\lim_{i\to\infty}V^{(i)}(\mathbf{s})=V(\mathbf{s})$, $\forall\mathbf{s}\in\mathcal{S}$. Thus, $h(\mathbf{s})=V(\mathbf{s})-V(\mathbf{s}_{\mathrm{ref}})$ satisfies (14) and $\bar{C}^\star=V(\mathbf{s}_{\mathrm{ref}})$. Functions V and h are (sometimes) called value function and relative value function, respectively. It is worth noting that any function h satisfying (14) is unique up to an additive factor, i.e., if h satisfies (14), so does $h+\alpha$, where α is any constant. The proposed RVIA is presented in Algorithm 1, where θ is a small constant for the RVIA termination criterion.

It is important to point out that the state space $\mathcal{S}$ and action space $\mathcal{A}$ grow exponentially in the number of sensors K, and thus, the complexity of the RVIA presented in Algorithm 1 grows exponentially in K. This is because the computational complexity for each iteration of the value iteration algorithm is $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$, where $|\mathcal{S}|$ is the number of states and $|\mathcal{A}|$ is the number of actions. Namely, finding an optimal policy is PSPACE-hard [46, Chap. 6]. Accordingly, finding an optimal policy $\pi^\star$ is practical only for small numbers of sensors. To this end, we next propose a low-complexity sub-optimal algorithm whose complexity increases only linearly in K.

IV. Low-Complexity Algorithm Design: Relax-Then-Truncate Approach

In this section, to handle massive IoT scenarios, we propose a low-complexity algorithm that provides a sub-optimal solution to problem (**P1**). The key observation is that the *per-slot* constraint (1) couples the actions $a_k(t)$, $k\in\mathcal{K}$, which results in the exponential complexity of finding an optimal policy for (**P1**), as explained in Section III. Therefore, we start by relaxing the per-slot constraint (1) into a time average constraint and subsequently model the *relaxed problem* as a *constrained MDP* (CMDP). The CMDP problem is then transformed into an unconstrained MDP problem through the Lagrangian approach [47]. The MDP problem decouples along the sensors and, therefore, for a fixed value of the Lagrange multiplier, we can find a per-sensor optimal policy. The optimal value of the Lagrange multiplier is found via bisection. This provides an optimal policy for the relaxed problem, called *optimal relaxed policy* hereinafter. Finally, we propose an online *truncation* procedure to ensure that the constraint (1) is satisfied at each slot. We remark that our optimality analysis in Section V shows that the proposed relax-then-truncate approach is *asymptotically optimal* as the number of sensors goes to infinity.

A. CMDP Formulation

We relax the constraint (1) and formulate the relaxed problem as a CMDP. To this end, we define the average number of command actions under a policy π_{R} as

$$\bar{J}_{\pi_{\mathrm{R}}}\triangleq\lim_{T\to\infty}\frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}_{\pi_{\mathrm{R}}}[a_k(t)], \tag{18}$$

and express the relaxed problem as

$$(\mathbf{P2})\quad \pi_{\mathrm{R}}^\star\in\arg\min_{\pi_{\mathrm{R}}}\bar{C}_{\pi_{\mathrm{R}}} \tag{19a}$$

$$\text{subject to } \bar{J}_{\pi_{\mathrm{R}}}\le\Gamma \tag{19b}$$

where $\Gamma\triangleq\frac{M}{K}$ is the normalized transmission budget.

We model (**P2**) as a CMDP defined by the tuple $\big(\mathcal{S},\mathcal{A}_{\mathrm{R}},\Pr(\mathbf{s}(t+1)|\mathbf{s}(t),\mathbf{a}(t)),c(\mathbf{s}(t),\mathbf{a}(t))\big)$, where the state space $\mathcal{S}$ and the relaxed action space $\mathcal{A}_{\mathrm{R}}$ were defined in Section II-E, and $\Pr(\mathbf{s}(t+1)|\mathbf{s}(t),\mathbf{a}(t))$ and $c(\mathbf{s}(t),\mathbf{a}(t))$ were defined in Section III-A. Note that the only difference between the CMDP tuple and the MDP tuple in Section III-A is in the action space ($\mathcal{A}_{\mathrm{R}}$ vs $\mathcal{A}$).

It is worth noting that any policy π that satisfies the per-slot transmission constraint (1) satisfies the time average transmission constraint (19b) in (**P2**). Thus, the average cost obtained by following policy $\pi^\star_{\mathrm{R}}$ is a *lower bound* on the average cost obtained under policy $\pi^\star$, i.e.,

$$\bar{C}_{\pi^\star_{\mathrm{R}}} \le \bar{C}_{\pi^\star}. \tag{20}$$

To solve the CMDP problem (**P2**), we introduce a Lagrange multiplier μ and define the Lagrangian associated with problem (**P2**) as

$$\mathcal{L}(\pi_{\mathrm{R}},\mu) \triangleq \lim_{T\to\infty} \frac{1}{NKT}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}_{\pi_{\mathrm{R}}}[c_k(t)+\mu a_k(t)] - \mu\frac{\Gamma}{N}. \tag{21}$$

For a given $\mu \ge 0$, we define the Lagrange dual function $\mathcal{L}^\star(\mu) = \min_{\pi_{\mathrm{R}}} \mathcal{L}(\pi_{\mathrm{R}},\mu)$. A policy that achieves $\mathcal{L}^\star(\mu)$ is called *μ-optimal*, denoted by $\pi^\star_{\mathrm{R},\mu}$, and it is a solution of the following (unconstrained) MDP problem

$$(\mathbf{P3}) \quad \pi^\star_{\mathrm{R},\mu} \in \arg\min_{\pi_{\mathrm{R}}} \mathcal{L}(\pi_{\mathrm{R}},\mu). \tag{22}$$

Since the dimension of the state space $\mathcal{S}$ is finite, the growth condition [47, Eq. 11.21] is satisfied. Moreover, the immediate cost function is bounded below, i.e., $c(\mathbf{s},\mathbf{a}) \ge 0$, $\forall \mathbf{a},\mathbf{s}$. Having these conditions satisfied, the optimal value of the CMDP problem (**P2**), $\bar{C}_{\pi^\star_{\mathrm{R}}}$, and the optimal value of the MDP problem (**P3**), $\mathcal{L}^\star(\mu)$, ensures the following relation [47, Corollary 12.2]

$$\bar{C}_{\pi^\star_{\mathrm{R}}} = \sup_{\mu\ge 0}\mathcal{L}^\star(\mu). \tag{23}$$

Therefore, an optimal policy for (**P2**) is found by a two-stage iterative algorithm: 1) for a given μ, we find a μ-optimal policy, and 2) we update μ in a direction that obtains $\bar{C}_{\pi^\star_{\mathrm{R}}}$ according to (23). These two steps are detailed in Sections IV-A.1 and IV-A.2, respectively.

1) An Optimal Policy for a Fixed Lagrange Multiplier: For a given μ, the problem of finding an optimal policy $\pi^\star_{\mathrm{R},\mu}$ in (**P3**) is *separable* across sensors $k \in \mathcal{K}$. Thus, (**P3**) can be decoupled into K per-sensor problems as follows. We express the Lagrangian in (21) equivalently as $\mathcal{L}(\pi_{\mathrm{R}},\mu) = \frac{1}{NK}\sum_{k=1}^{K}\mathcal{L}_k(\pi_k,\mu) - \mu\frac{\Gamma}{N}$, where $\mathcal{L}_k(\pi_k,\mu)$ is defined as

$$\mathcal{L}_k(\pi_k,\mu) \triangleq \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi_k}[c_k(t)+\mu a_k(t)], \quad k=1,\dots,K, \tag{24}$$

where the per-sensor policy π_k was defined in Section II-E.3. Thus, finding an optimal policy $\pi^\star_{\mathrm{R},\mu}$ reduces to finding K per-sensor optimal policies, denoted by $\pi^\star_{\mathrm{R},\mu,k}$, $k=1,\dots,K$, as

$$(\mathbf{P4}) \quad \pi^\star_{\mathrm{R},\mu,k} \in \arg\min_{\pi_k} \mathcal{L}_k(\pi_k,\mu), \quad k=1,\dots,K. \tag{25}$$

Each sub-problem (**P4**), for a particular k, can be modeled as an (unconstrained) MDP problem. Particularly, we define the MDP model associated with sensor k as the tuple $(\mathcal{S}_k, \mathcal{A}_k, \Pr(s_k(t+1)|s_k(t),a_k(t)), c_k(s_k(t),a_k(t))+\mu a_k(t))$, where the per-sensor state space $\mathcal{S}_k$ and the per-sensor action space $\mathcal{A}_k$ were defined in Section II-E, the per-sensor state transition probabilities $\Pr(s_k(t+1)|s_k(t),a_k(t))$ are calculated as in (10), and the cost of taking action $a_k(t)$ in state $s_k(t)$ is $c_k(s_k(t),a_k(t))+\mu a_k(t)$, where $c_k(s_k(t),a_k(t))$ is defined in Section III-A.

*Proposition 3: The per-sensor MDP formulated for (**P4**) is communicating, i.e., for every pair of states $s, s' \in \mathcal{S}_k$, there exists a stationary policy under which s' is accessible from s.*

Proof: The proof is presented in Appendix VII-B. □

By Proposition 3 and Proposition 2, and rewriting the Bellman's equation in (14) for the per-sensor MDP formulation, we have

$$\begin{aligned}&\mathcal{L}^\star_k(\mu) + h_{\mathrm{R},\mu,k}(s)\\ &\quad= \min_{a\in\mathcal{A}_k}\Bigg[c_k(s,a)+\mu a + \sum_{s'\in\mathcal{S}_k}\Pr(s'|s,a)h_{\mathrm{R},\mu,k}(s')\Bigg],\\ &\quad\ s\in\mathcal{S}_k,\end{aligned} \tag{26}$$

where $\mathcal{L}^\star_k(\mu) \triangleq \min_{\pi_k}\mathcal{L}_k(\pi_k,\mu)$. In addition, an optimal policy in state $s \in \mathcal{S}_k$ is given by

$$\begin{aligned}\pi^\star_{\mathrm{R},\mu,k}(s) \in \arg\min_{a\in\mathcal{A}_k}\Bigg[&c_k(s,a)+\mu a\\ &+\sum_{s'\in\mathcal{S}_k}\Pr(s'|s,a)h_{\mathrm{R},\mu,k}(s')\Bigg], \quad s\in\mathcal{S}_k.\end{aligned} \tag{27}$$

By turning (26) into an iterative procedure, $h_{\mathrm{R},\mu,k}(s)$ and consequently $\pi^\star_{\mathrm{R},\mu,k}(s)$, $s\in\mathcal{S}_k$, are obtained iteratively. Particularly, at each iteration $i=0,1,\dots$, we have

$$\begin{aligned}V^{(i+1)}_{\mathrm{R},\mu,k}(s) = \min_{a\in\mathcal{A}_k}\ &[c_k(s,a)+\mu a\\ &+\sum_{s'\in\mathcal{S}_k}\Pr(s'|s,a)h^{(i)}_{\mathrm{R},\mu,k}(s')],\end{aligned} \tag{28}$$

$$h^{(i+1)}_{\mathrm{R},\mu,k}(s) = V^{(i+1)}_{\mathrm{R},\mu,k}(s) - V^{(i+1)}_{\mathrm{R},\mu,k}(s_{\mathrm{ref}}), \tag{29}$$

where $s_{\mathrm{ref}}\in\mathcal{S}_k$ is an arbitrary reference state. For any initialization $V^{(0)}_{\mathrm{R},\mu,k}(s)$, the sequences $\{h^{(i)}_{\mathrm{R},\mu,k}(s)\}_{i=1,2,\dots}$ and $\{V^{(i)}_{\mathrm{R},\mu,k}(s)\}_{i=1,2,\dots}$ converge, i.e., $\lim_{i\to\infty}h^{(i)}_{\mathrm{R},\mu,k}(s) = h_{\mathrm{R},\mu,k}(s)$ $\lim_{i\to\infty}V^{(i)}_{\mathrm{R},\mu,k}(s) = V_{\mathrm{R},\mu,k}(s)$, $\forall s \in \mathcal{S}_k$. Thus, $h_{\mathrm{R},\mu,k}(s) = V_{\mathrm{R},\mu,k}(s) - V_{\mathrm{R},\mu,k}(s_{\mathrm{ref}})$ satisfies (26) and $\mathcal{L}^\star_k(\mu) = V_{\mathrm{R},\mu,k}(s_{\mathrm{ref}})$. The proposed RVIA is presented in Algorithm 2 (Lines 17–35).

Next, we give insight to optimal policies by studying the structure of $\pi^\star_{\mathrm{R},\mu,k}$ obtained by the proposed RVIA. Besides, this inherent structure may be exploited to design structural-aware RVIA that further reduces the computational complexity of the RVIA (see e.g., [7], [48]). We first prove that $V_{\mathrm{R},\mu,k}(s)$ has monotonic properties and then exploit them to prove that a per-sensor optimal policy has a threshold-based structure with respect to the AoI.

Lemma 1: Function $V_{\mathrm{R},\mu,k}$ is non-decreasing with respect to the AoI, i.e., for any two states $\underline{s} = (r,b,\underline{\Delta}) \in \mathcal{S}_k$ and $s = (r,b,\Delta)\in\mathcal{S}_k$ with $\underline{\Delta}\ge\Delta$, we have $V(\underline{s}) \ge V(s)$.

Proof: The proof is presented in Appendix VII-C. □

Theorem 1: For the case where the link from sensor k to the edge node is perfect (i.e., $\xi_k = 1$), a per-sensor optimal policy

$\pi^\star_{\mathrm{R},\mu,k}$ *obtained by RVIA has a threshold-based structure with respect to the AoI, i.e., if* $\pi^\star_{\mathrm{R},\mu,k}(s)=1$ *in state* $s=(r,b,\Delta)$, *then for all states* $\underline{s}=(r,b,\underline{\Delta})$, $\underline{\Delta}\geq\Delta$, *an optimal action is also* $\pi^\star_{\mathrm{R},\mu,k}(\underline{s})=1$.

Proof: The proof is presented in Appendix VII-D. □

2) Determination of the Optimal Lagrange Multiplier: Recall that the cost function associated with the per-sensor MDP formulation (established for (**P4**)) is defined as $c_k(s_k(t),a_k(t))+\mu a_k(t)$. Hence, by increasing μ, the cost of taking action $a_k(t)=1$ increases, and thus, the edge node tends to use the command action less. More precisely, $\bar{C}_{\pi^\star_{\mathrm{R},\mu}}$ and $\mathcal{L}(\pi^\star_{\mathrm{R},\mu},\mu)$ are increasing in μ, whereas $\bar{J}_{\pi^\star_{\mathrm{R},\mu}}$ is decreasing in μ [49, Lemma 3.1]. Therefore, we are interested in the smallest value of the Lagrange multiplier such that policy $\pi^\star_{\mathrm{R},\mu}$ satisfies the time average transmission constraint (19b). Formally, we define the optimal Lagrange multiplier as [49]

$$\mu^* \triangleq \inf\Big\{\mu\geq 0 \mid \bar{J}_{\pi^\star_{\mathrm{R},\mu}}\leq\Gamma\Big\}, \tag{30}$$

where $\bar{J}_{\pi^\star_{\mathrm{R},\mu}}$ is the average number of command actions under policy $\pi^\star_{\mathrm{R},\mu}$, which is calculated using (18). From (18) and the fact that (**P3**) is decoupled into K per-sensor problems (**P4**), $\bar{J}_{\pi^\star_{\mathrm{R},\mu}}$ is calculated as $\bar{J}_{\pi^\star_{\mathrm{R},\mu}}=\frac{1}{K}\sum_{k=1}^{K}\bar{J}_{\pi^\star_{\mathrm{R},\mu,k}}$, where $\bar{J}_{\pi^\star_{\mathrm{R},\mu,k}}$ denotes the per-sensor time average number of command actions under the per-sensor policy $\pi^\star_{\mathrm{R},\mu,k}$, which is defined as

$$\bar{J}_{\pi^\star_{\mathrm{R},\mu,k}} \triangleq \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi^\star_{\mathrm{R},\mu,k}}[a_k(t)]. \tag{31}$$

Thus, (30) is rewritten as

$$\mu^*=\inf\left\{\mu\geq 0:\sum_{k=1}^{K}\bar{J}_{\pi^\star_{\mathrm{R},\mu,k}}\leq K\Gamma\right\}. \tag{32}$$

We now characterize an optimal relaxed policy $\pi^\star_{\mathrm{R}}$ for (**P2**). If the average number of command actions obtained by $\pi^\star_{\mathrm{R},\mu^*,k}$ satisfies $\frac{1}{K}\sum_{k=1}^{K}\bar{J}_{\pi^\star_{\mathrm{R},\mu^*,k}}=\Gamma$, then, $\pi^\star_{\mathrm{R},\mu^*,k}$, $k\in\mathcal{K}$, form an optimal policy for (**P2**), i.e., $\pi^\star_{\mathrm{R}}=\pi^\star_{\mathrm{R},\mu^*}$. Otherwise, $\pi^\star_{\mathrm{R}}$ is a mixture of two deterministic policies $\pi^\star_{\mathrm{R},\mu^{*-}}$ and $\pi^\star_{\mathrm{R},\mu^{*+}}$, which are defined by [49, Theorem 4.4]

$$\pi^\star_{\mathrm{R},\mu^{*-}} \triangleq \lim_{\mu\to\mu^{*-}}\pi^\star_{\mathrm{R},\mu} \text{ and } \pi^\star_{\mathrm{R},\mu^{*+}} \triangleq \lim_{\mu\to\mu^{*+}}\pi^\star_{\mathrm{R},\mu}, \tag{33}$$

and is written symbolically as $\pi^\star_{\mathrm{R}}\triangleq\eta\pi^\star_{\mathrm{R},\mu^{*-}}+(1-\eta)\pi^\star_{\mathrm{R},\mu^{*+}}$, where η is the mixing factor. This mixed policy is a stationary randomized policy where the action at each state s is $\pi^\star_{\mathrm{R},\mu^{*-}}(\mathbf{s})$ with probability η and $\pi^\star_{\mathrm{R},\mu^{*+}}(\mathbf{s})$ with probability $1-\eta$, where η is obtained[7] such that $\bar{J}_{\pi^\star_{\mathrm{R}}}=\Gamma$.

To search for μ^* as defined in (30), we apply the bisection method that exploits the monotonicity of $\bar{J}_{\pi^\star_{\mathrm{R},\mu}}$ with respect to μ. Particularly, if $\frac{1}{K}\sum_{k=1}^{K}\bar{J}_{\pi^\star_{\mathrm{R},\mu,k}}\leq\Gamma$ for $\mu=0$, then the constraint (19b) is inactive, and an optimal policy for (**P2**) is $\pi^\star_{\mathrm{R},0}$. Otherwise, we apply an iterative update procedure until $|\mu^+-\mu^-|<\epsilon$ and $\frac{1}{K}\sum_{k=1}^{K}\bar{J}_{\pi^\star_{\mathrm{R},\mu,k}}\leq\Gamma$ are satisfied. Details are expressed in Algorithm 2, where ϵ is a small constant for the bisection termination criterion.

Algorithm 2 Policy Design for the CMDP Problem (**P2**) via RVIA and Bisection Search

```
Initialize Set $\mu \leftarrow 0$, $\mu^- \leftarrow 0$, $\mu^+$ as a large positive number, and determine a small $\epsilon > 0$
RVIA($\mu$)            ▷ run function RVIA for input $\mu = 0$
if $\bar{J}_{\pi^\star_{\mathrm{R},\mu}} \leq \Gamma$ then
    $\pi^\star_{\mathrm{R}} = \pi^\star_{\mathrm{R},\mu}$
else
    while $|\mu^+ - \mu^-| > \epsilon$ do
     RVIA($\frac{\mu^+ + \mu^-}{2}$)   ▷ run function RVIA for $\mu = \frac{\mu^+ + \mu^-}{2}$
      if $\bar{J}_{\pi^\star_{\mathrm{R},\mu}} \geq \Gamma$ then $\mu^- \leftarrow \mu$ else $\mu^+ \leftarrow \mu$
    end while
    $\mu^* \leftarrow 1/2(\mu^- + \mu^+)$, $\mu^{*-} \leftarrow \mu^-$, and $\mu^{*+} \leftarrow \mu^+$
    if $\bar{J}_{\pi^\star_{\mathrm{R},\mu}} = \Gamma$ then
        $\pi^\star_{\mathrm{R}} = \pi^\star_{\mathrm{R},\mu^*}$
    else
        $\pi^\star_{\mathrm{R}} = \eta\pi^\star_{\mathrm{R},\mu^{*-}} + (1-\eta)\pi^\star_{\mathrm{R},\mu^{*+}}$
    end if
end if

function RVIA($\mu$)    ▷ find optimal policies $\pi^\star_{\mathrm{R},\mu,k}$ using RVIA for fixed $\mu$
    Initialize $V_{\mathrm{R},\mu,k}(s) \leftarrow 0$, $h_{\mathrm{R},\mu,k}(s) \leftarrow 0$, $\forall k, \forall s$,
    determine $s_{\mathrm{ref}} \in \mathcal{S}_k$ and a small $\theta > 0$
    for $k = 1, \ldots, K$ do
        repeat
            for $s \in \mathcal{S}_k$ do
                $V_{\mathrm{tmp}}(s) \leftarrow \min_{a\in\mathcal{A}_k}[c_k(s,a) + \mu a$
                $+\sum_{s'\in\mathcal{S}_k}\Pr(s' \mid s,a)h_{\mathrm{R},\mu,k}(s')]$
            end for
          $\delta \leftarrow \max_{s\in\mathcal{S}_k}(V_{\mathrm{tmp}}(s) - V_{\mathrm{R},\mu,k}(s)) -$
              $\min_{s\in\mathcal{S}_k}(V_{\mathrm{tmp}}(s) - V_{\mathrm{R},\mu,k}(s))$, for all $s \in \mathcal{S}_k$
          $V_{\mathrm{R},\mu,k}(s) \leftarrow V_{\mathrm{tmp}}(s)$, for all $s \in \mathcal{S}_k$
          $h_{\mathrm{R},\mu,k}(s) \leftarrow V_{\mathrm{R},\mu,k}(s) - V_{\mathrm{R},\mu,k}(s_{\mathrm{ref}})$, for all $s \in \mathcal{S}_k$
        until $\delta < \theta$
        $\pi^\star_{\mathrm{R},\mu,k}(s) = \arg\min_{a\in\mathcal{A}_k}[c_k(s,a) + \mu a +$
        $\sum_{s'\in\mathcal{S}_k}\Pr(s' \mid s,a)\,h_{\mathrm{R},\mu,k}(s')]$, for all $s \in \mathcal{S}_k$
    end for
    Output: per-sensor optimal policies $\pi^\star_{\mathrm{R},\mu,k}$, $k \in \mathcal{K}$
end function
```

Remark 2: It is worth noting that the complexity of finding an optimal relaxed policy $\pi^\star_{\mathrm{R}}$ increases linearly in the number of sensors K, whereas the complexity of finding an optimal policy $\pi^\star$ grows exponentially in K. Consider a scenario with $K=100$ sensors, $N=7$ users, $\Delta^{\max}=64$, and $B_k=15$. The size of the state space $\mathcal{S}$ is $|\mathcal{S}|=(8\times16\times64)^{100}=2^{1300}\approx10^{400}$. However, the per-sensor state space size is $|\mathcal{S}_k|=8\times16\times64=2^{13}\approx10^{6}$.

B. Truncation Procedure

Recall that there is no guarantee that the per-slot constraint (1) is satisfied under an optimal relaxed policy $\pi^\star_{\mathrm{R}}$. Here, we propose the following truncation procedure to satisfy

[7] There is no closed-form expression for η [50]. Therefore, we can numerically search for such $\eta\in[0,1]$.

Algorithm 3 Truncation Procedure

```
Input Optimal relaxed policy $\pi^\star_{\mathrm{R}}$
for each slot $t = 1, 2, 3, \ldots$ do
    Construct the set $\mathcal{X}(t)$ based on $\pi^\star_{\mathrm{R}}$
    if $|\mathcal{X}(t)| \le M$ then
        $a_k(t) = 1$, for all $k \in \mathcal{X}(t)$
    else
        Select $M$ sensors from $\mathcal{X}(t)$ randomly (uniform)
        and command them
    end if
end for
```

the constraint (1) at each slot. For slot t, we define a set $\mathcal{X}(t) \triangleq \{k \mid a_k(t) = 1, k \in \mathcal{K}\} \subseteq \mathcal{K}$ that represents the set of sensors that are commanded under $\pi^\star_{\mathrm{R}}$. The truncation procedure divides into two cases: 1) if $|\mathcal{X}(t)| \le M$, the edge node commands all the sensors in $\mathcal{X}(t)$, and 2) otherwise, the edge node selects M sensors from the set $\mathcal{X}(t)$ *randomly* according to the discrete uniform distribution and commands them to send status updates. The online truncation procedure is presented in Algorithm 3.

V. Asymptotic Optimality of the Proposed Relax-Then-Truncate Approach

In this section, we analyze the optimality of the proposed relax-then-truncate policy – denoted by $\tilde{\pi}$ hereinafter – developed in Section IV. We first find an upper bound for the difference between the average cost obtained by the policy $\tilde{\pi}$ and the average cost obtained by an optimal policy $\pi^\star$. Then, we present two lemmas that are used to show that the relax-then-truncate approach is *asymptotically optimal* as the number of sensors goes to infinity.

Theorem 2: The difference between the average cost obtained by the relax-then-truncate policy $\tilde{\pi}$ and the average cost obtained by an optimal policy $\pi^\star$ is upper bounded as

$$\bar{C}_{\tilde{\pi}} - \bar{C}_{\pi^\star} \le \frac{\Delta^{\max}}{M} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \underbrace{\mathbb{E}_{\pi^\star_{\mathrm{R}}}\left[\left||\mathcal{X}(t)| - \mathbb{E}_{\pi^\star_{\mathrm{R}}}\left[|\mathcal{X}(t)|\right]\right|\right]}_{\triangleq \mathrm{MAD}(|\mathcal{X}(t)|)}, \tag{34}$$

where $\mathrm{MAD}(\cdot)$ *denotes the Mean Absolute Deviation.*

Proof: The proof is presented in Appendix VII-E. □

We next present two lemmas that will subsequently be used in Theorem 3 to prove the asymptotic optimality of the relax-then-truncate approach.

Lemma 2: For a random variable X that follows a normal distribution with mean ν and variance σ^2, i.e., $X \sim \mathcal{N}(\nu, \sigma^2)$, the mean absolute deviation is given as $\mathrm{MAD}(X) = \sqrt{\frac{2}{\pi}}\sigma$.

Proof: The proof is presented in Appendix VII-F. □

Lemma 3: When $K \to \infty$, by following the policy $\pi^\star_{\mathrm{R}}$, we have $\mathrm{MAD}\left(\frac{|\mathcal{X}(t)|}{\sqrt{K}}\right) \le 1$.

Proof: The proof is presented in Appendix VII-G. □

Theorem 3: For a fixed $\Gamma = M/K$, the relax-then-truncate policy $\tilde{\pi}$ is asymptotically optimal with respect to the number of sensors, i.e., $\lim_{K\to\infty}(\bar{C}_{\tilde{\pi}} - \bar{C}_{\pi^\star}) = 0$.

Proof: The proof is presented in Appendix VII-H. □

VI. Simulation Results

In this section, we provide simulation results to evaluate the performance of the low-complexity relax-then-truncate approach developed in Section IV and illustrate the structure of per-sensor optimal policies attained by the RVIA in Algorithm 2.

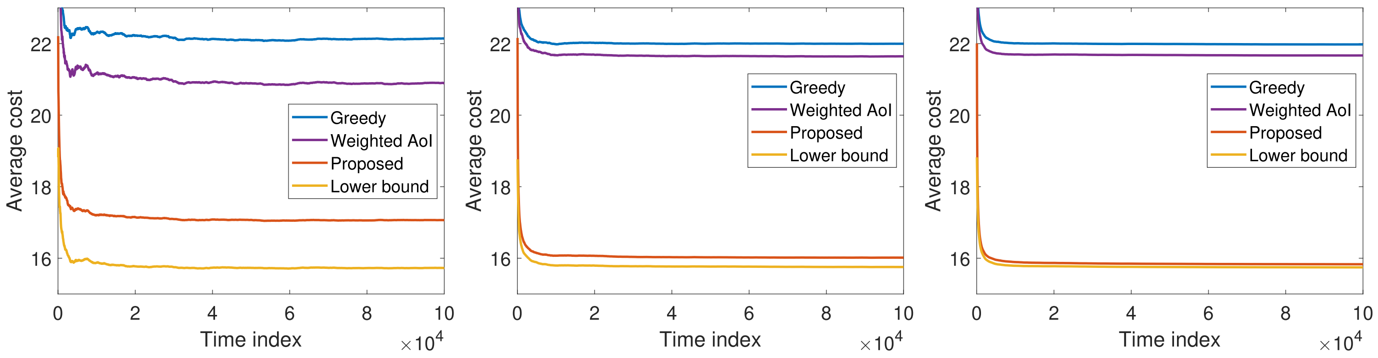
A. Performance of the Proposed Low-Complexity Relax-Then-Truncate Approach

We consider an IoT network with $N = 3$ users, where in each slot, user n requests a status of f_k with probability $p_{k,n} = 0.6$. The battery capacity of each sensor is set to $B_k = 7$ units of energy, the transmit success probability is set to $\xi_k = 0.8$, and the AoI upper-bound is set to $\Delta^{\max} = 64$. Each sensor is assigned an energy harvesting rate λ_k from the set $\{0.01, 0.02, \ldots, 0.1\}$ in the following sequential order: sensors $1, 11, 21, \ldots$ are assigned the energy harvesting rate 0.01, sensors $2, 12, 22, \ldots$ are assigned the energy harvesting rate 0.02 etc.

We compare the performance of the proposed relax-then-truncate policy with a *(request-aware) greedy* policy, a *weighted AoI* policy, and a lower bound. In the greedy policy, the edge node commands at most M sensors with the largest AoI from the set $\mathcal{W}(t) \triangleq \{k \mid r_k(t) \ge 1, k \in \mathcal{K}\}$, i.e., the set of sensors whose measurements are requested by at least one user. In the weighted AoI policy, the edge node commands M sensors with the highest value of $r_k(t)\Delta_k(t)$; randomization is used in case of a tie. The lower bound is obtained by following $\pi^\star_{\mathrm{R}}$ (see (20)).

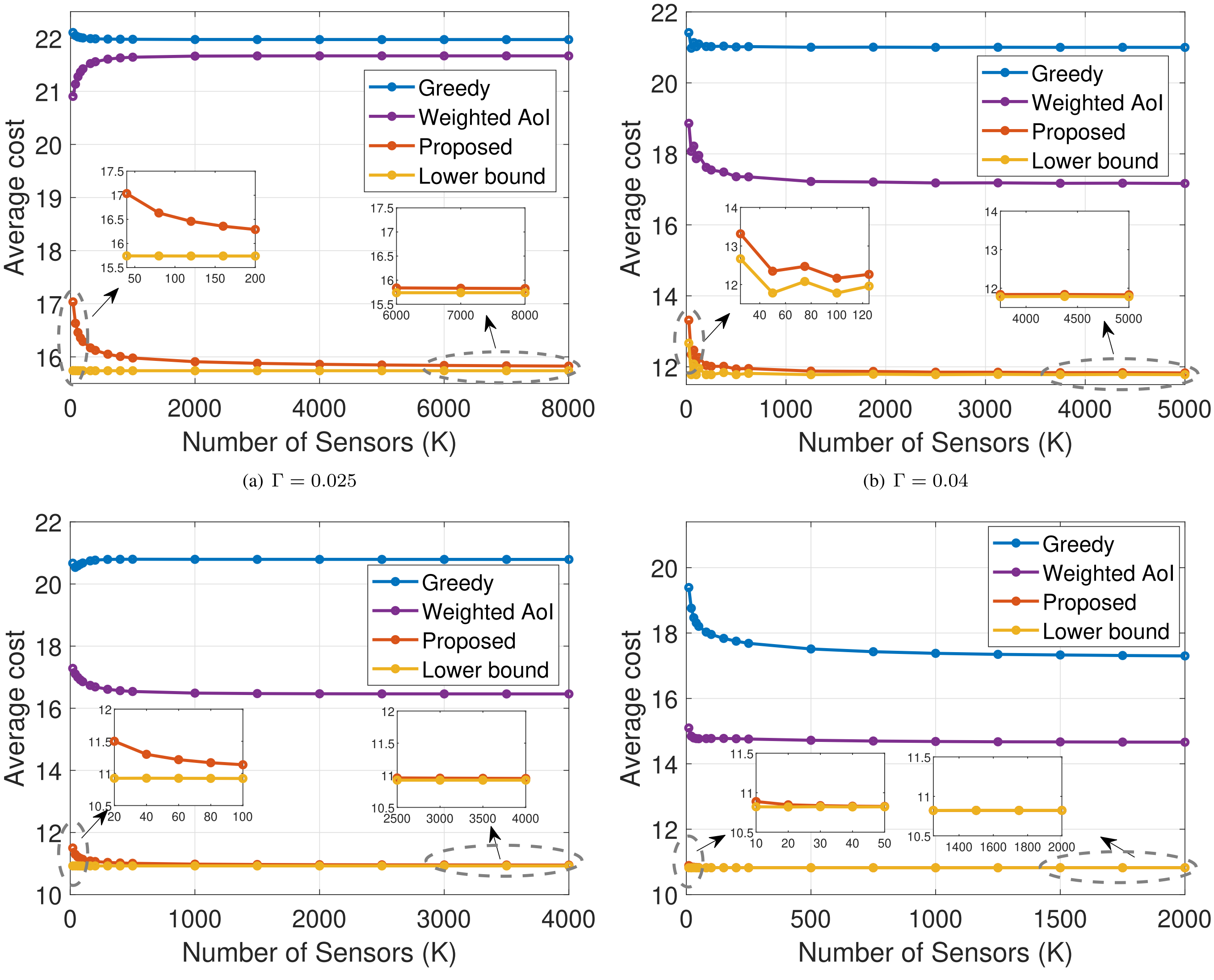
Fig. 2 depicts the performance of the relax-then-truncate algorithm over time for different numbers of sensors K with a fixed normalized transmission budget $\Gamma = 0.025$. As shown, the proposed algorithm reduces the average cost by approximately 37.5% compared to the greedy policy. Furthermore, the gap between the proposed policy and the lower bound is in general small and decreases as K increases. The proposed policy approaches the lower bound for large K, which validates the asymptotic optimality of the proposed algorithm as proved in Theorem 3.

Fig. 3 depicts the performance of the relax-then-truncate algorithm with respect to the number of sensors K for different values of Γ. The results are obtained by averaging each algorithm over 10 episodes where each episode takes 5×10^6 slots. Due to asymptotic optimality of the proposed method, the gap between the proposed policy and the lower bound diminishes for large values of K. Furthermore, the proposed policy performs near-optimally for moderate numbers of sensors, which is important for practical use cases. Figs. 3(a)–(d) show that, as Γ increases, the proposed policy converges to the optimal performance faster. This is because, by increasing Γ, the proportion of the sensors that can be commanded at each slot increases, and consequently, the proportion of truncated sensors (i.e., the sensors that are not commanded under $\tilde{\pi}$ compared to $\pi^\star_{\mathrm{R}}$) decreases. Besides, the weighted AoI policy outperforms the greedy policy because, in addition to the AoI, the exact number of requests are considered in its action selection.



(a) $M = 1$ and $K = 40$ (b) $M = 20$ and $K = 800$ (c) $M = 200$ and $K = 8000$

Fig. 2. Performance of the proposed relax-then-truncate algorithm in terms of average cost (i.e., average on-demand AoI over all the sensors and users) over time for different values of the number of sensors K with a fixed normalized transmission budget $\Gamma = 0.025$.

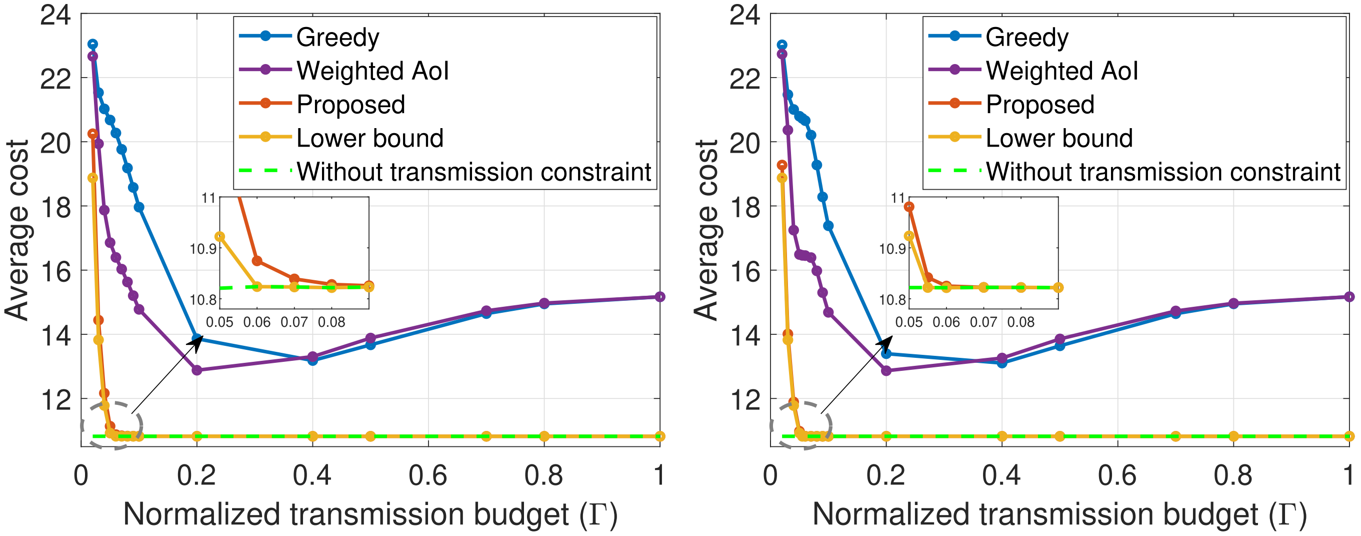


(a) $\Gamma = 0.025$ (b) $\Gamma = 0.04$

(c) $\Gamma = 0.05$ (d) $\Gamma = 0.1$

Fig. 3. Performance of the proposed relax-then-truncate approach in terms of average cost with respect to the number of sensors K for different values of Γ.

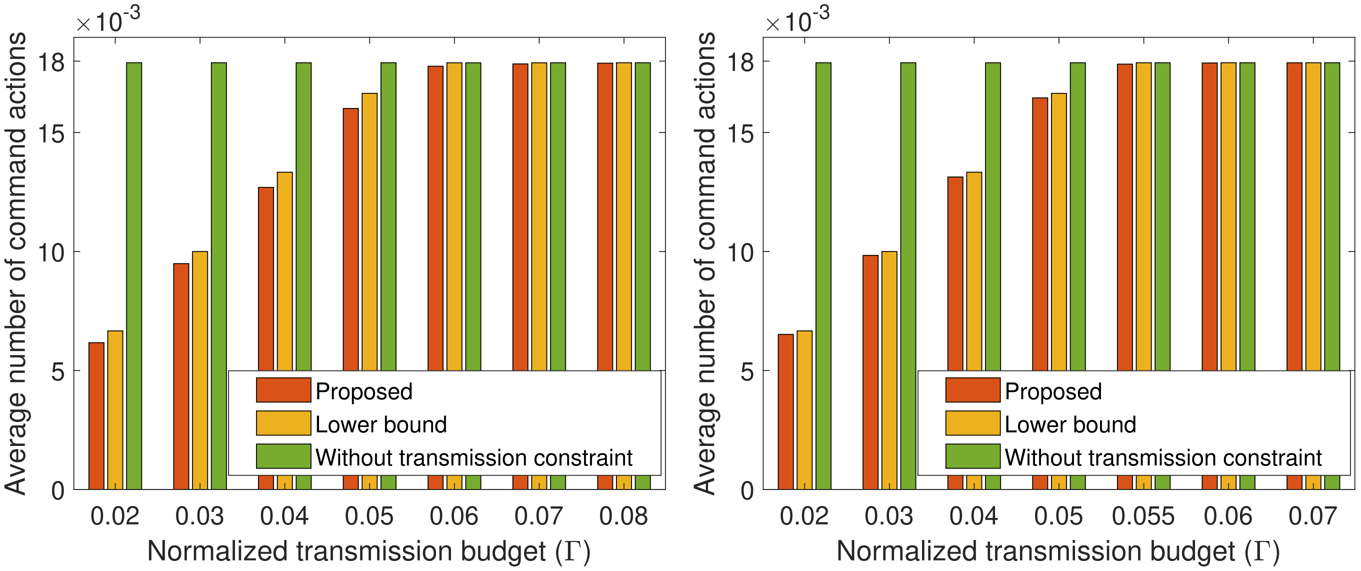
Fig. 4 and Fig. 5 illustrate the average cost and the average number of command actions, respectively, with respect to Γ. The performance of an optimal policy for the case without any transmission constraint (i.e., $M = K$) is depicted as a benchmark [34]. As shown in Fig. 4, the average cost for the proposed algorithm decreases as Γ increases. This is because, for fixed K, the transmission budget M increases by increasing Γ, and thus, more sensors can be commanded at each slot so that the users are served with fresh measurements more often. Interestingly, from a certain point onward, increasing



(a) $K = 100$

(b) $K = 1000$

Fig. 4. Performance of the proposed algorithm in terms of average cost with respect to Γ.



(a) $K = 100$

(b) $K = 1000$

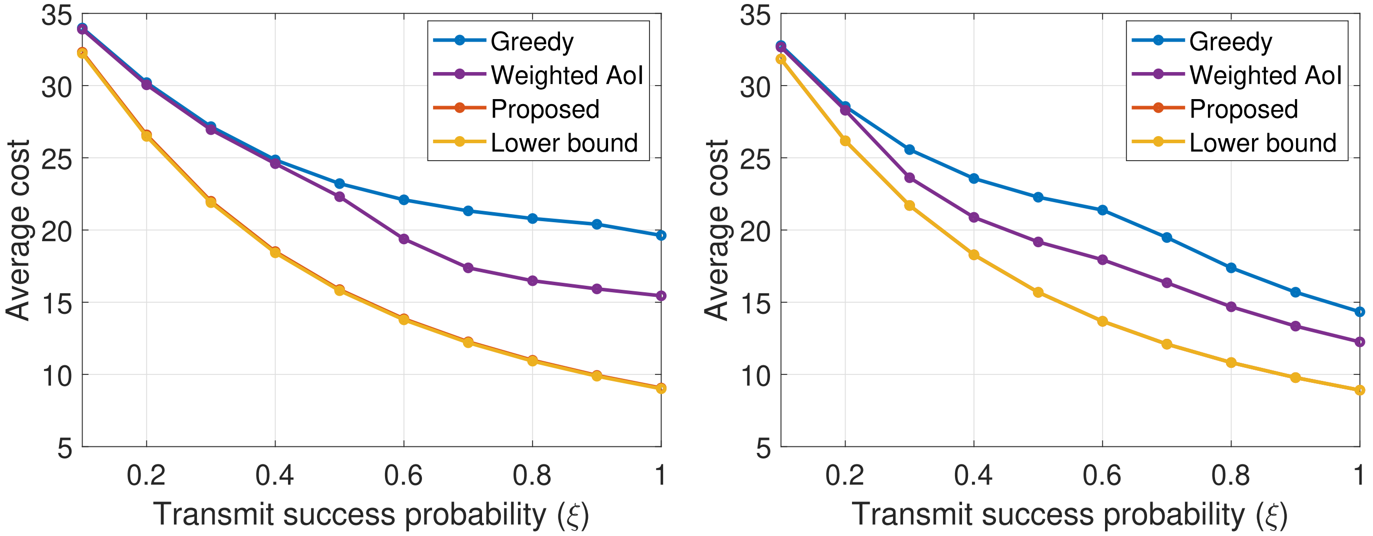
Fig. 5. Average number of command actions with respect to Γ.

Γ does not decrease the average cost. This is because the average number of command actions stops increasing as shown in Fig. 5, i.e., the constraint (19b) becomes inactive as the edge node has more transmission budget than needed. In these cases, the limited availability of energy at the EH sensors becomes a dominant factor in restraining the transmission of fresh status updates. To exemplify, for the case where $K = 1000$, the network reaches its maximum performance when $M = 60$, and therefore, increasing the transmission budget (e.g., bandwidth) further is not effective. Such observation is important for practical applications because increasing M would incur additional cost in the networks.

Fig. 6 shows the average cost with respect to the transmit success probability ξ ($\xi_k = \xi$, $\forall k$) for different values of Γ, where $K = 1000$. As shown in Fig. 6, the average cost decreases as ξ increases. This is because, by increasing ξ, the communication link from the sensor to the edge node becomes more reliable, thus increasing the probability that the edge node successfully receives fresh status update packets from the sensors.

B. Structural Properties of Per-Sensor Deterministic Policies for the Relaxed Problem

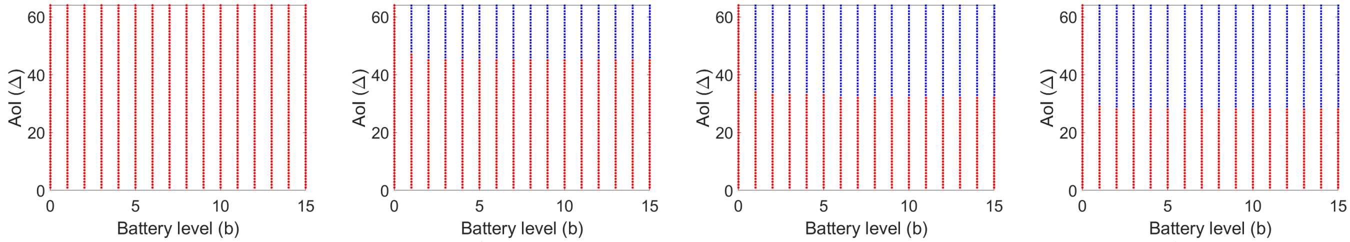
Here, we consider a setup with $K = 400$ identical sensors with battery capacity $B_k = 15$ units of energy. We analyze the structural properties of a per-sensor policy obtained by Algorithm 2 for a particular sensor k, i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}$, and investigate the effect of the transmission budget M, energy harvesting rate λ_k, and request probability $p_{k,n}$. Fig. 7 illustrates the structure of $\pi^{\star}_{\mathrm{R},\mu^*,k}$, where each point represents a potential per-sensor state as a three-tuple $s = (r, b, \Delta)$. For each such state, a blue point indicates that the optimal action is to command the sensor (i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}(s) = 1$), whereas a red point means not



(a) $\Gamma = 0.05$

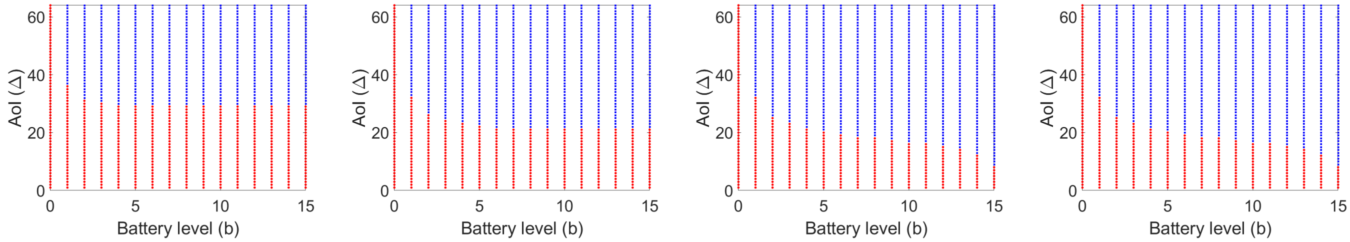
(b) $\Gamma = 0.1$

Fig. 6. Performance of the proposed algorithm in terms of average cost with respect to ξ.



(a) $r = 0$ (b) $r = 1$ (c) $r = 2$ (d) $r = 3$

Fig. 7. Structure of an optimal policy for sensor k (i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}$) for each state $s = \{r, b, \Delta\}$, where $M = 10$, $\lambda_k = 0.06$, $\xi_k = 1$, and $p_{k,n} = 0.2$. Red: no command; blue: command.



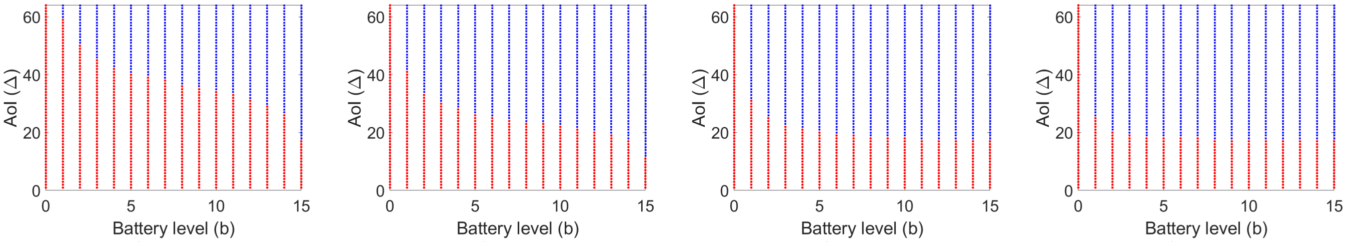
(a) $M = 15$ (b) $M = 20$ (c) $M = 25$ (d) $M = 30$

Fig. 8. Structure of an optimal policy for sensor k (i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}$) in states $s = \{1, b, \Delta\}$ for different numbers of transmission budget M, where $\lambda_k = 0.06$, $\xi_k = 1$, and $p_{k,n} = 0.2$.

to command (i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}(s) = 0$). The set of the blue points is referred to as the *command region* hereinafter.

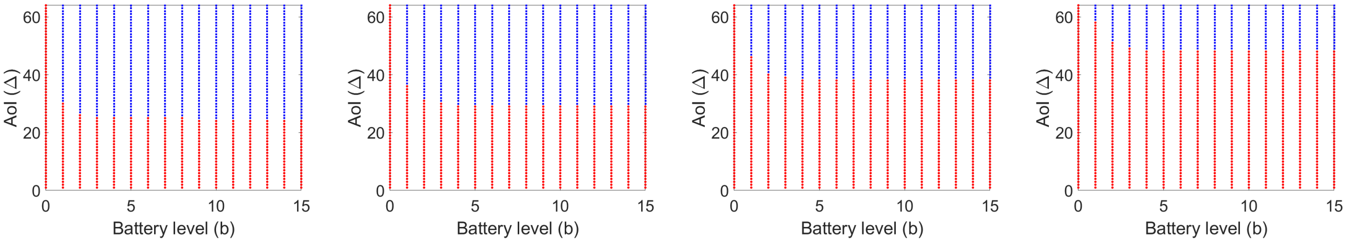
Fig. 7 shows that $\pi^{\star}_{\mathrm{R},\mu^*,k}$ has *threshold-based* structure with respect to the number of requests r, battery level b, and AoI Δ. Consider a state $s = (1, 5, 50)$ in which $\pi^{\star}_{\mathrm{R},\mu^*,k}(s) = 1$; then, by the threshold-based structure, $\pi^{\star}_{\mathrm{R},\mu^*,k}(\underline{s}) = 1$ for all states $\underline{s} = (r, b, \Delta)$, $r \geq 1$, $b \geq 5$, $\Delta \geq 50$. Furthermore, Fig. 7 manifests the impact of considering the on-demand AoI (instead of conventional AoI) as the objective cost. Namely, since the cost function (6) is (linearly) increasing with $r_k(t)$, the edge node has more incentive to command a sensor that is associated with a large number of requests. As expected, if there are no requests for f_k (i.e., $r_k = 0$), the optimal action is not to command the sensor, regardless of the battery level and AoI, i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}(0, b, \Delta) = 0$. This way the sensor conserves (and possibly recharges) its battery to be able to respond to upcoming users' requests.

Fig. 8, Fig. 9, and Fig. 10 depict the action under $\pi^{\star}_{\mathrm{R},\mu^*,k}$ in each state $s = \{1, b, \Delta\}$ for different values of the transmission budget M, energy harvesting rate λ_k, and request probability $p_{k,n}$, respectively. It is inferred from Figs. 8(a)–(d) that the command region enlarges as M increases, because the edge node can command more sensors at each slot. Further, from a certain point onward ($M \geq 25$), the command region does not expand anymore, because the sensors' energy limitation restrains the number of commands for new status updates. A comparison in Figs. 9(a)–(d) shows that the command region is enlarged by increasing λ_k; This is because when



(a) $\lambda_k = 0.04$ (b) $\lambda_k = 0.06$ (c) $\lambda_k = 0.08$ (d) $\lambda_k = 0.10$

Fig. 9. Structure of an optimal policy for sensor k (i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}$) in states $s = \{1, b, \Delta\}$ for different values of energy harvesting rate λ_k, where $M = 30$, $\xi_k = 1$ and $p_{k,n} = 0.4$.



(a) $p_{k,n} = 0.1$ (b) $p_{k,n} = 0.2$ (c) $p_{k,n} = 0.4$ (d) $p_{k,n} = 0.6$

Fig. 10. Structure of an optimal policy for sensor k (i.e., $\pi^{\star}_{\mathrm{R},\mu^*,k}$) in states $s = \{1, b, \Delta\}$ for different values of request probability $p_{k,n}$, where $M = 15$, $\lambda_k = 0.06$, and $\xi_k = 1$.

a sensor harvests energy more often, it can send updates more often. Finally, as shown in Figs. 10(a)–(d), when the sensors are requested more often (i.e., $p_{k,n}$ increases), the command region shrinks; the edge node commands the sensor less to save its energy for the future requests.

VII. Conclusion

We investigated on-demand AoI minimization problem in a resource-constrained IoT network, where multiple users make on-demand requests to a cache-enabled edge node to send status updates about various random processes, each monitored by an EH sensor. We first modeled the problem as an MDP and proposed an iterative algorithm that obtains an optimal policy. Since the complexity of finding an optimal policy increases exponentially in the number of sensors, we developed a low-complexity relax-then-truncate algorithm and then analytically showed that it is asymptotically optimal as the number of sensors goes to infinity. Numerical results illustrated that the relax-then-truncate algorithm significantly reduces the average cost (i.e., average on-demand AoI over all sensors and users) compared to a request-aware greedy policy and a weighted AoI policy, and performs close to the optimal solution for moderate numbers of sensors.

Appendix

A. Proof of Proposition 1

Proof: For any state $\mathbf{s} = (s_1, \ldots, s_K)$, where $s_k = (r_k, b_k, \Delta_k)$, $k = 1, \ldots, K$, we define the request vector $\mathbf{r} = (r_1, \ldots, r_K)$, the battery vector $\mathbf{b} = (b_1, \ldots, b_K)$, and the age vector $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)$. Recall that at most M sensors can send a fresh status update at each slot. Thus, any state whose age vector has more than M identical entries with values strictly less than $\Delta^{\max}$ is a transient state. We consider two non-transient states $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_{\mathrm{c}}$ and show that $\mathbf{s}' \triangleq (\mathbf{r}', \mathbf{b}', \boldsymbol{\Delta}')$ is accessible from $\mathbf{s} \triangleq (\mathbf{r}, \mathbf{b}, \boldsymbol{\Delta})$ under a stationary randomized policy π in which, at each state $\mathbf{s}$, the edge node randomly selects an action $\mathbf{a} \in \mathcal{A}$ according to the discrete uniform distribution, i.e., $\pi(\mathbf{a}|\mathbf{s}) = \frac{1}{|\mathcal{A}|}$. Let δ denote the largest element of the age vector $\boldsymbol{\Delta}'$ (i.e., $\max_k \Delta'_k = \delta$). Let $\mathbf{e}_i$ denote a unit vector of length K having a single 1 at the ith entry and all other entries 0. Let $\mathbf{e}_0$ denote a zero vector (i.e., all entries are 0) of length K. We define a vector $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,K})$ with elements $a_{i,k} = \mathbb{1}_{\{\Delta'_k = i,\ \Delta'_k < \Delta^{\max}\}}$. First, since the requests processes are independent from other variables in the system (e.g., actions), a state with a request vector $\mathbf{r}'$ is accessible from any other state. Second, realizing the actions $\mathbf{e}_1$ for $(b_1 - b'_1)^+$ slots, $\mathbf{e}_2$ for $(b_2 - b'_2)^+$ slots, $\ldots$, $\mathbf{e}_K$ for $(b_K - b'_K)^+$ slots, and $\mathbf{e}_0$ for $\tau = \max_k |b'_k - b_k| - \sum_k (b_k - b'_k)^+$ slots, the system reaches a state whose battery vector is $\mathbf{b}'$ with a positive probability (w.p.p.). Note that, regardless of the actions happening next, the system reaches a state whose battery vector is still $\mathbf{b}'$ w.p.p. Third, realizing the consecutive actions $\mathbf{a}_\delta$, $\mathbf{a}_{\delta-1}$, $\ldots$, $\mathbf{a}_1$ leads the system reach a state whose age vector is $\boldsymbol{\Delta}'$ w.p.p. In summary, the system reaches a state with request vector $\mathbf{r}'$, age vector $\boldsymbol{\Delta}'$, and battery vector $\mathbf{b}'$ w.p.p.. Thus, $\mathbf{s}'$ is accessible from $\mathbf{s}$. □

B. Proof of Proposition 3

Proof: We consider two arbitrary states $s, s' \in \mathcal{S}_k$ and show that $s' = (r', b', \Delta')$ is accessible from $s = (r, b, \Delta)$ under a (per-sensor) stationary randomized policy π_k in which, at each state s, the edge node randomly selects an action $a \in \mathcal{A}_k = \{0, 1\}$ according to the discrete uniform distribution, i.e., $\pi_k(0|s) = \pi_k(1|s) = 1/2$. For the case where $b' \geq b$, realizing the action $a = 0$ for $\tau = b' - b + 1$ consecutive slots leads to state $(r', b' + 1, \min\{\Delta + \tau, \Delta^{\max}\})$ w.p.p.; then the action $a = 1$ leads to state $(r', b', 1)$ w.p.p., and subsequently action $a = 0$ for $\Delta' - 1$ consecutive slots leads to state $s' = (r', b', \Delta')$ w.p.p. Similarly, for the case where $b' < b$, the action $a = 1$ for $\tau = b - b'$ consecutive slots leads to state $(r', b', 1)$ w.p.p., and subsequently $a = 0$ for $\Delta' - 1$ consecutive slots leads to state $s' = (r', b', \Delta')$ w.p.p. □

C. Proof of Lemma 1

Proof: For brevity, we drop the unnecessary subscripts, e.g., $V_{\mathrm{R},\mu,k}$ is simply shown by V. We consider $\underline{s} = (r, b, \underline{\Delta})$ and $s = (r, b, \Delta)$ with $\underline{\Delta} \geq \Delta$ and prove that $V(s) \leq V(\underline{s})$. Since the sequence $\{V^{(i)}(s)\}_{i=1,2,\ldots}$ converges to $V(s)$ for any initialization, it suffices to prove that $V^{(i)}(\underline{s}) \geq V^{(i)}(s)$, $\forall i$, which is shown using mathematical induction. The initial values are selected arbitrarily, e.g., $V^{(0)}(s) = 0$ and $V^{(0)}(\underline{s}) = 0$, hence, $V^{(i)}(\underline{s}) \geq V^{(i)}(s)$ holds for $i = 0$. Assume that $V^{(i)}(\underline{s}) \geq V^{(i)}(s)$ for some i; we need to prove that $V^{(i+1)}(\underline{s}) \geq V^{(i+1)}(s)$. We define $Q^{(i+1)}(s, a) \triangleq c_k(s, a) + \mu a + \sum_{s' \in \mathcal{S}_k} \Pr(s'|s, a) h^{(i)}(s')$, $s \in \mathcal{S}_k, a \in \mathcal{A}_k$. Thus, $V^{(i+1)}(s) = \min_{a \in \mathcal{A}_k} Q^{(i+1)}(s, a)$ (see (28)). Let us denote an optimal action in state s at iteration $i = 1, 2, \ldots$ by $\pi^{(i)}(s)$, which is given by $\pi^{(i)}(s) = \arg\min_{a \in \mathcal{A}_k} Q^{(i)}(s, a)$. We have

$$
\begin{aligned}
&V^{(i+1)}(s) - V^{(i+1)}(\underline{s}) \\
&\quad= \min_{a \in \mathcal{A}_k} Q^{(i+1)}(s, a) - \min_{a \in \mathcal{A}_k} Q^{(i+1)}(\underline{s}, a) \\
&\quad= Q^{(i+1)}(s, \pi^{(i+1)}(s)) - Q^{(i+1)}(\underline{s}, \pi^{(i+1)}(\underline{s})) \\
&\quad\overset{(a)}{\leq} Q^{(i+1)}(s, \pi^{(i+1)}(\underline{s})) - Q^{(i+1)}(\underline{s}, \pi^{(i+1)}(\underline{s})),
\end{aligned}
$$

where (a) follows from the fact that taking action $\pi^{(i+1)}(\underline{s})$ in state s is not necessarily optimal. We show that $Q^{(i+1)}(s, \pi^{(i+1)}(\underline{s})) - Q^{(i+1)}(\underline{s}, \pi^{(i+1)}(\underline{s})) \leq 0$ for all possible actions $\pi^{(i+1)}(\underline{s}) \in \{0, 1\}$. The proof is presented for the case where $\pi^{(i+1)}(\underline{s}) = 1$ and $b \geq 1$; the proof follows similarly for the other three cases, i.e., $\pi^{(i+1)}(\underline{s}) = 0$ and $b < B_k$, $\pi^{(i+1)}(\underline{s}) = 0$ and $b = B_k$, and $\pi^{(i+1)}(\underline{s}) = 1$ and $b = 0$. We have the relations in (35), shown at the bottom of the page, where in step (a) we used (10)–(13), step (b) follows from the assumption $\Delta \leq \underline{\Delta}$, and step (c) follows from the induction assumption. □

D. Proof of Theorem 1

Proof: For brevity, we drop the unnecessary subscripts, e.g., $V_{\mathrm{R},\mu,k}$ is simply shown by V. Let us define $Q(s, a) \triangleq c_k(s, a) + \mu a + \sum_{s' \in \mathcal{S}_k} \Pr(s'|s, a) h(s')$. Thus, $V(s) = \min_{a \in \mathcal{A}_k} Q(s, a)$. Proving that $\pi^\star$ has a threshold-based structure with respect to the AoI is equivalent to showing the following: if the optimal action in state $s = (r, b, \Delta)$ is $\pi^\star(s) = 1$, i.e., $Q(s, 1) - Q(s, 0) \leq 0$, then for all states $\underline{s} = (r, b, \underline{\Delta})$ with $\underline{\Delta} \geq \Delta$ the optimal action is also $\pi^\star(\underline{s}) = 1$, i.e., $Q(\underline{s}, 1) - Q(\underline{s}, 0) \leq 0$. This is equivalent to showing that $Q(\underline{s}, 1) - Q(\underline{s}, 0) \leq Q(s, 1) - Q(s, 0)$. The proof is presented for the case where $1 \leq b < B_k$; the proof follows similarly for the other two cases, i.e., $b = B_k$ and $b = 0$. We have the relations in (36), shown at the bottom of the page, where step (a) follows from the assumption $\Delta \leq \underline{\Delta}$ and step (b) follows from Lemma 1. □

E. Proof of Theorem 2

Proof: Let $\mathcal{T}(t) \subset \mathcal{X}(t)$ denote the set of *truncated* sensors at slot t, i.e., the sensors that are not commanded under the relax-then-truncate policy $\tilde{\pi}$, given that they are commanded under policy $\pi^\star_{\mathrm{R}}$. By the truncation procedure, if $|\mathcal{X}(t)| > M$, M sensors are chosen randomly (uniform) from the set $\mathcal{X}(t)$ and commanded (i.e., $|\mathcal{X}(t)| - M$ sensors

$$
\begin{aligned}
&Q^{(i+1)}(s, 1) - Q^{(i+1)}(\underline{s}, 1) \\
&\quad= c_k(s, 1) + \sum_{s' \in \mathcal{S}_k} \Pr(s'|s, 1) V^{(i)}(s') - c_k(\underline{s}, 1) - \sum_{\underline{s}' \in \mathcal{S}_k} \Pr(\underline{s}'|\underline{s}, 1) V^{(i)}(\underline{s}') \\
&\quad\overset{(a)}{=} r(1 - \xi_k) \underbrace{\left(\min\{\Delta + 1, \Delta^{\max}\} - \min\{\underline{\Delta} + 1, \Delta^{\max}\}\right)}_{(b) \leq 0} + \sum_{n=0}^{N} \sum_{l=0}^{1} \Pr(r' = n) \left(l\lambda_k + (1 - l)(1 - \lambda_k)\right) (1 - \xi_k) \\
&\qquad \times \underbrace{\left(V^{(i)}(n, b + l - 1, \min\{\Delta + 1, \Delta^{\max}\}) - V^{(i)}(n, b + l - 1, \min\{\underline{\Delta} + 1, \Delta^{\max}\})\right)}_{(c) \leq 0} \leq 0,
\end{aligned}
\tag{35}
$$

$$
\begin{aligned}
&Q(s, 1) - Q(\underline{s}, 1) - Q(s, 0) + Q(\underline{s}, 0) \\
&\quad= c_k(s, 1) + \sum_{s' \in \mathcal{S}_k} \Pr(s'|s, 1) V(s') - c_k(\underline{s}, 1) - \sum_{\underline{s}' \in \mathcal{S}_k} \Pr(\underline{s}'|\underline{s}, 1) V(\underline{s}') - c_k(s, 0) \\
&\qquad - \sum_{s' \in \mathcal{S}_k} \Pr(s'|s, 0) V(s') + c_k(\underline{s}, 0) + \sum_{\underline{s}' \in \mathcal{S}_k} \Pr(\underline{s}'|\underline{s}, 0) V(\underline{s}') = r \underbrace{\left(\min\{\underline{\Delta} + 1, \Delta^{\max}\} - \min\{\Delta + 1, \Delta^{\max}\}\right)}_{(a) \geq 0} \\
&\qquad + \sum_{n=0}^{N} \sum_{l=0}^{1} \Pr(r' = n)(l\lambda_k + (1 - l)(1 - \lambda_k)) \underbrace{\left(V(n, b + l, \min\{\underline{\Delta} + 1, \Delta^{\max}\}) - V(n, b + l, \min\{\Delta + 1, \Delta^{\max}\})\right)}_{(b) \geq 0} \geq 0,
\end{aligned}
\tag{36}
$$

are not commanded). Hence, the probability that sensor k belongs to $\mathcal{T}(t)$ is $\mathbb{1}_{\{|\mathcal{X}(t)|>M|\}}\left(\frac{|\mathcal{X}(t)|-M}{|\mathcal{X}(t)|}\right)$. At each slot, the additional per-sensor cost under $\tilde{\pi}$ compared to $\pi_{\mathrm{R}}^{\star}$ is no more than $N\Delta^{\max}$ (see (6)). Therefore, the expected additional cost over all sensors under $\tilde{\pi}$ compared to $\pi_{\mathrm{R}}^{\star}$ is upper bounded by

$$\sum_{k=1}^{K}\underbrace{\mathbb{1}_{\{\mathcal{X}(t)>M\}}\frac{|\mathcal{X}(t)|-M}{|\mathcal{X}(t)|}}_{\Pr(k\in\mathcal{T}(t))}N\Delta^{\max}$$
$$= NK\Delta^{\max}\frac{(|\mathcal{X}(t)|-M)^{+}}{|\mathcal{X}(t)|}, \qquad (37)$$

where $(\cdot)^{+}\triangleq\max\{0,\cdot\}$.

We introduce the following (penalized) strategy $\hat{\pi}_{\mathrm{R}}$: at each slot, command the sensors based on $\pi_{\mathrm{R}}^{\star}$ but add a penalty $NK\Delta^{\max}\frac{(|\mathcal{X}(t)|-M)^{+}}{|\mathcal{X}(t)|}$ to the cost over all sensors (see (37)). It is clear that the average cost obtained under $\hat{\pi}_{\mathrm{R}}$ is not less than that obtained by $\tilde{\pi}$, i.e., $\bar{C}_{\tilde{\pi}}\le\bar{C}_{\hat{\pi}_{\mathrm{R}}}$. Also, recall from (20) that the average cost obtained under policy $\pi_{\mathrm{R}}^{\star}$ is a lower bound for the average cost obtained by an optimal policy $\pi^{\star}$, i.e., $\bar{C}_{\pi_{\mathrm{R}}^{\star}}\le\bar{C}_{\pi^{\star}}$. Moreover, policy $\tilde{\pi}$ is a sub-optimal solution for (**P1**), i.e., $\bar{C}_{\pi^{\star}}\le\bar{C}_{\tilde{\pi}}$. Therefore, we have

$$\bar{C}_{\pi_{\mathrm{R}}^{\star}}\le\bar{C}_{\pi^{\star}}\le\bar{C}_{\tilde{\pi}}\le\bar{C}_{\hat{\pi}_{\mathrm{R}}}. \qquad (38)$$

Using (38), the difference between the average cost obtained by the proposed relax-then-truncate policy $\tilde{\pi}$ and the average cost obtained by an optimal policy $\pi^{\star}$ is upper bounded as

$$\begin{aligned}
\bar{C}_{\tilde{\pi}}-\bar{C}_{\pi^{\star}} &\overset{(a)}{\le} \bar{C}_{\hat{\pi}_{\mathrm{R}}}-\bar{C}_{\pi_{\mathrm{R}}^{\star}}\\
&= \lim_{T\to\infty}\frac{1}{NKT}\sum_{t=1}^{T}\mathbb{E}_{\pi_{\mathrm{R}}^{\star}}\left[NK\Delta^{\max}\frac{(|\mathcal{X}(t)|-M)^{+}}{|\mathcal{X}(t)|}\right]\\
&\overset{(b)}{\le} \frac{\Delta^{\max}}{M}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi_{\mathrm{R}}^{\star}}\left[(|\mathcal{X}(t)|-M)^{+}\right]\\
&\overset{(c)}{\le} \frac{\Delta^{\max}}{M}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi_{\mathrm{R}}^{\star}}\left[(|\mathcal{X}(t)|-\mathbb{E}_{\pi_{\mathrm{R}}^{\star}}[|\mathcal{X}(t)|])^{+}\right]\\
&\overset{(d)}{\le} \frac{\Delta^{\max}}{M}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi_{\mathrm{R}}^{\star}}\Big[\big||\mathcal{X}(t)|-\mathbb{E}_{\pi_{\mathrm{R}}^{\star}}[|\mathcal{X}(t)|]\big|\Big]\\
&= \frac{\Delta^{\max}}{M}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{MAD}(|\mathcal{X}(t)|), \qquad (39)
\end{aligned}$$

where (a) follows from (38), (b) follows from $\frac{(|\mathcal{X}(t)|-M)^{+}}{|\mathcal{X}(t)|}\le\frac{(|\mathcal{X}(t)|-M)^{+}}{M}$, (c) follows from $\mathbb{E}_{\pi_{\mathrm{R}}^{\star}}[|\mathcal{X}(t)|]\le M$, for sufficiently large t, and (d) follows from $(\cdot)^{+}\le|\cdot|$. □

F. Proof of Lemma 2

$$\begin{aligned}
\mathrm{MAD}(X) &= \mathbb{E}[|X-\nu|]=\int_{-\infty}^{\infty}|x-\nu|\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\nu}{\sigma}\right)^{2}}\mathrm{d}x\\
&= \int_{-\infty}^{\nu}(\nu-x)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\nu}{\sigma}\right)^{2}}\mathrm{d}x\\
&\quad+\int_{\nu}^{\infty}(x-\nu)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\nu}{\sigma}\right)^{2}}\mathrm{d}x\\
&= \sqrt{\frac{2}{\pi}}\sigma\int_{0}^{\infty}ye^{-\frac{1}{2}y^{2}}\mathrm{d}y=\sqrt{\frac{2}{\pi}}\sigma. \qquad (40)
\end{aligned}$$

G. Proof of Lemma 3

Proof: The cardinality of set $\mathcal{X}(t)$ (i.e., the set of sensors that are commanded under $\pi_{\mathrm{R}}^{\star}$) can be written as $|\mathcal{X}(t)|=\sum_{k=1}^{K}a_k(t)$, where $a_k(t)\in\{0,1\}$, $k\in\mathcal{K}$, are K independent binary random variables. Let $\omega_k(t)$ be the probability that sensor k is commanded at slot t under policy $\pi_{\mathrm{R}}^{\star}$, i.e., $\omega_k(t)\triangleq\Pr(a_k(t)=1)$. We define a random variable $Z(t)\triangleq\frac{|\mathcal{X}(t)|-\sum_k\omega_k(t)}{\sqrt{\sum_k\omega_k(t)(1-\omega_k(t))}}$. We have

$$\begin{aligned}
\mathrm{MAD}\left(Z(t)\right) &= \mathrm{MAD}\left(\frac{|\mathcal{X}(t)|-\sum_k\omega_k(t)}{\sqrt{\sum_k\omega_k(t)(1-\omega_k(t))}}\right)\\
&\overset{(a)}{=} \mathrm{MAD}\left(\frac{|\mathcal{X}(t)|}{\sqrt{\sum_k\omega_k(t)(1-\omega_k(t))}}\right)\\
&\overset{(b)}{\ge} \mathrm{MAD}\left(\frac{|\mathcal{X}(t)|}{\sqrt{K/4}}\right)\ge\mathrm{MAD}\left(\frac{|\mathcal{X}(t)|}{\sqrt{K}}\right),
\end{aligned} \qquad (41)$$

where (a) follows because the MAD does not change by adding a constant to all values of the variable (similar to variance) and (b) follows from $\sum_{k=1}^{K}\omega_k(t)(1-\omega_k(t))\le\frac{K}{4}$.

By the Lyapunov central limit theorem [51, Theorem 27.3], $Z(t)$ converges in distribution to a standard normal distribution, i.e., $Z(t)\sim\mathcal{N}(0,1)$, as K goes to infinity. Thus, we have

$$\lim_{K\to\infty}\mathrm{MAD}\left(\frac{|\mathcal{X}(t)|}{\sqrt{K}}\right)\overset{(a)}{\le}\lim_{K\to\infty}\mathrm{MAD}(Z(t))\overset{(b)}{=}\sqrt{\frac{2}{\pi}}\le 1, \qquad (42)$$

where (a) follows from (41) and (b) follows from Lemma 2. □

H. Proof of Theorem 3

Proof: We have

$$\begin{aligned}
&\lim_{K\to\infty}\left(\bar{C}_{\tilde{\pi}}-\bar{C}_{\pi^{\star}}\right)\\
&\overset{(a)}{\le}\lim_{K\to\infty}\left[\frac{\Delta^{\max}}{\Gamma\sqrt{K}}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{MAD}\left(\frac{|\mathcal{X}(t)|}{\sqrt{K}}\right)\right]\\
&\overset{(b)}{\le}\lim_{K\to\infty}\frac{\Delta^{\max}}{\Gamma\sqrt{K}}=0, \qquad (43)
\end{aligned}$$

where (a) follows from Theorem 2 and $M=\Gamma K$, and (b) follows from Lemma 3. □

References

[1] M. Hatami, M. Leinonen, Z. Chen, N. Pappas, and M. Codreanu, "Asymptotically optimal on-demand AoI minimization in energy harvesting IoT networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Espoo, Finland, Jun. 2022, pp. 922–927.

[2] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 2731–2735.

[3] Y. Sun, I. Kadota, R. Talak, and E. Modiano, "Age of information: A new metric for information freshness," *Synth. Lectures Commun. Netw.*, vol. 12, no. 2, pp. 1–224, Dec. 2019.

[4] D. Niyato, D. I. Kim, P. Wang, and L. Song, "A novel caching mechanism for Internet of Things (IoT) sensing service with energy harvesting," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[5] M. Hatami, M. Leinonen, and M. Codreanu, "AoI minimization in status update control with energy harvesting sensors," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8335–8351, Dec. 2021.

[6] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.

[7] Y.-P. Hsu, E. Modiano, and L. Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Trans. Mobile Comput.*, vol. 19, no. 12, pp. 2903–2915, Dec. 2020.

[8] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2637–2650, Dec. 2018.

[9] A. Maatouk, S. Kriouile, M. Assad, and A. Ephremides, "On the optimality of the Whittle's index policy for minimizing the age of information," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1263–1277, Feb. 2021.

[10] S. Kriouile, M. Assaad, and A. Maatouk, "On the global optimality of Whittle's index policy for minimizing the age of information," 2021, *arXiv:2102.02528*.

[11] E. T. Ceran, D. Gunduz, and A. Gyorgy, "Average age of information with hybrid ARQ under a resource constraint," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1900–1913, Mar. 2019.

[12] E. T. Ceran, D. Gunduz, and A. Gyorgy, "A reinforcement learning approach to age of information in multi-user networks with HARQ," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1412–1426, May 2021.

[13] H. Tang, J. Wang, L. Song, and J. Song, "Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 854–868, Mar. 2020.

[14] A. Zakeri, M. Moltafet, M. Leinonen, and M. Codreanu, "Minimizing AoI in resource-constrained multi-source relaying systems with stochastic arrivals," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–6.

[15] X. Wu, X. Li, J. Li, P. C. Ching, and H. V. Poor, "Deep reinforcement learning for lot networks: Age of information and energy cost tradeoff," in *Proc. IEEE Global Telecommun. Conf.*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

[16] X. Wu, X. Li, J. Li, P. C. Ching, V. C. M. Leung, and H. V. Poor, "Caching transient content for IoT sensing: Multi-agent soft actor-critic," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5886–5901, Sep. 2021.

[17] B. T. Bacinoglu, E. T. Ceran, and E. Uysal-Biyikoglu, "Age of information under energy replenishment constraints," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2015, pp. 25–31.

[18] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 193–204, Mar. 2018.

[19] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 534–556, Jan. 2020.

[20] E. Tugce Ceran, D. Gunduz, and A. Gyorgy, "Learning to minimize age of information over an unreliable channel with energy harvesting," 2021, *arXiv:2106.16037*.

[21] S. Feng and J. Yang, "Age of information minimization for an energy harvesting source with updating erasures: Without and with feedback," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5091–5105, Aug. 2021.

[22] C. Tunc and S. Panwar, "Optimal transmission policies for energy harvesting age of information systems with battery recovery," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 2012–2016.

[23] S. Leng and A. Yener, "Age of information minimization for an energy harvesting cognitive radio," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 2, pp. 427–439, Jun. 2019.

[24] Z. Chen, N. Pappas, E. Bjornson, and E. G. Larsson, "Optimizing information freshness in a multiple access channel with heterogeneous devices," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 456–470, 2021.

[25] G. Stamatakis, N. Pappas, and A. Traganitis, "Control of status updates for energy harvesting devices that monitor processes with alarms," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[26] E. Gindullina, L. Badia, and D. Gunduz, "Age-of-information with information source diversity in an energy harvesting system," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 3, pp. 1529–1540, Sep. 2021.

[27] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "AoI-optimal joint sampling and updating for wireless powered communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 14110–14115, Nov. 2020.

[28] M. Sheikhi and V. Hakami, "AoI-aware status update control for an energy harvesting source over an uplink mmWave channel," in *Proc. 7th Int. Conf. Signal Process. Intell. Syst. (ICSPIS)*, Tehran, Iran, Dec. 2021, pp. 1–6.

[29] A. Jaiswal, A. Chattopadhyay, and A. Varma, "Age-of-information minimization via opportunistic sampling by an energy harvesting source," 2022, *arXiv:2201.02787*.

[30] N. Zhao, C. Xu, S. Zhang, Y. Xie, X. Wang, and H. Sun, "Status update for correlated energy harvesting sensors: A deep reinforcement learning approach," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nanjing, China, Oct. 2020, pp. 170–175.

[31] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "A reinforcement learning framework for optimizing age of information in RF-powered communication systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4747–4760, Aug. 2020.

[32] A. K. Aydin and N. Akar, "Energy management for age of information control in solar-powered IoT end devices," *Wireless Netw.*, vol. 27, no. 5, pp. 3165–3178, Jul. 2021.

[33] M. Hatami, M. Jahandideh, M. Leinonen, and M. Codreanu, "Age-aware status update control for energy harvesting IoT sensors via reinforcement learning," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, London, U.K., Aug. 2020, pp. 1–10.

[34] M. Hatami, M. Leinonen, and M. Codreanu, "Minimizing average on-demand AoI in an IoT network with energy harvesting sensors," in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Lucca, Italy, Sep. 2021, pp. 1–6.

[35] B. Yin *et al.*, "Only those requested count: Proactive scheduling policies for minimizing effective age-of-information," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Paris, France, Apr. 2019, pp. 109–117.

[36] F. Li, Y. Sang, Z. Liu, B. Li, H. Wu, and B. Ji, "Waiting but not aging: Optimizing information freshness under the pull model," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 465–478, Feb. 2021.

[37] J. Holm *et al.*, "Freshness on demand: Optimizing age of information for the query process," in *Proc. IEEE Int. Conf. Commun.*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.

[38] F. Chiariotti *et al.*, "Query age of information: Freshness in pull-based communication," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 1606–1622, Mar. 2022.

[39] N. Michelusi, K. Stamatiou, and M. Zorzi, "Transmission policies for energy harvesting sensors with time-correlated energy supply," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2988–3001, Jul. 2013.

[40] N. Pappas, Z. Chen, and M. Hatami, "Average AoI of cached status updates for a process monitored by an energy harvesting sensor," in *Proc. 54th Annu. Conf. Inf. Sci. Syst. (CISS)*, Princeton, NJ, USA, Mar. 2020, pp. 1–5.

[41] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Orlando, FL, USA, Jun. 2015, pp. 3008–3012.

[42] E. T. Ceran, D. Gunduz, and A. Gyorgy, "Reinforcement learning to minimize age of information with an energy harvesting sensor with HARQ and sensing cost," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Paris, France, Apr. 2019, pp. 656–661.

[43] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Using erasure feedback for online timely updating with an energy harvesting sensor," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Orlando, FL, USA, Jul. 2019, pp. 607–611.

[44] Y. H. Wang, "On the number of successes in independent trials," *Statist. Sinica*, vol. 3, no. 2, pp. 295–312, 1993.

[45] D. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 2, 3rd ed. Nashua, NH, USA: Athena Scientific, 2007.

[46] J. Gittins, K. Glazebrook, and R. Weber, *Multi-Armed Bandit Allocation Indices*. Hoboken, NJ, USA: Wiley, 2011.

[47] E. Altman, *Constrained Markov Decision Processes*, vol. 7. Boca Raton, FL, USA: CRC Press, 1999.

[48] G. Yao, A. Bedewy, and N. B. Shroff, "Age-optimal low-power status update over time-correlated fading channel," *IEEE Trans. Mobile Comput.*, early access, Mar. 16, 2022, doi: 10.1109/TMC.2022.3160050.

[49] F. J. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint," *J. Math. Anal. Appl.*, vol. 112, no. 1, pp. 236–252, 1985.

[50] D.-J. Ma, A. Makowski, and A. Shwartz, "Estimation and optimal control for constrained Markov chains," in *Proc. 25th IEEE Conf. Decis. Control*, Athens, Greece, Dec. 1986, pp. 994–999.

[51] P. Billingsley, *Probability and Measure*, 3rd ed. Hoboken, NJ, USA: Wiley, 1995.

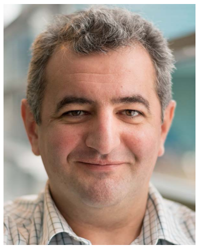
Mohammad Hatami (Graduate Student Member, IEEE) received the M.Sc. (Tech.) degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2015. He is currently pursuing the Ph.D. degree with the Centre for Wireless Communications, University of Oulu, Finland. His current research interests include information freshness in wireless networks, caching, machine learning for wireless applications, and system design.

Markus Leinonen (Member, IEEE) received the B.Sc. (Tech.) and M.Sc. (Tech.) degrees in electrical engineering and the D.Sc. (Tech.) degree in communications engineering from the University of Oulu, Finland, in 2010, 2011, and 2018, respectively. In 2010, he joined the Centre for Wireless Communications, University of Oulu. He received an Academy of Finland Post-Doctoral Researcher position from 2021 to 2024. In 2013, he was a Guest Researcher with the Technical University of Munich, Germany. In 2020, he was a Visiting Post-Doctoral Researcher with the University of California San Diego (UCSD). He has published over 50 journal and conference papers, a book, and a book chapter in the areas of wireless communications, mathematical optimization, and signal and information processing. His current research interests include time-critical networking and sparsity-aware signal processing for wireless communications. The research covers optimization and analysis of information freshness in status update systems and design of sparse signal recovery methods for channel estimation, user activity detection, and signal detection in wireless communication systems. He served as the Local Arrangements Chair for the IEEE SPAWC 2022. He was appointed as an Exemplary Reviewer of IEEE TRANSACTIONS ON COMMUNICATIONS in 2020.

Zheng Chen (Member, IEEE) received the B.Sc. degree from the Huazhong University of Science and Technology (HUST), China, in 2011, and the M.Sc. and Ph.D. degrees from the CentraleSupélec, Université Paris-Saclay, France, in 2013 and 2016, respectively.

From June 2015 to November 2015, she was a Visiting Scholar at the Singapore University of Technology and Design (SUTD), Singapore. Since January 2017, she has been with Linköping University, Sweden. She is currently an Assistant Professor with the Department of Electrical Engineering, Linköping University. Her main research interests include wireless communications, distributed intelligent systems, and complex networks. She was a recipient of the 2020 IEEE Communications Society Young Author Best Paper Award. She served as the Workshop Co-Chair for the IEEE GLOBECOM Workshop on Wireless Communications for Distributed Intelligence in 2021 and 2022. She was selected as an Exemplary Reviewer for IEEE COMMUNICATIONS LETTERS in 2016, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2017, and IEEE TRANSACTIONS ON COMMUNICATIONS in 2019.

Nikolaos Pappas (Senior Member, IEEE) received the B.Sc. degree in computer science, the M.Sc. degree in computer science, the B.Sc. degree in mathematics, and the Ph.D. degree in computer science from the University of Crete, Greece, in 2005, 2007, 2012, and 2012, respectively. From 2005 to 2012, he was a Graduate Research Assistant with the Telecommunications and Networks Laboratory, Institute of Computer Science, Foundation for Research and Technology—Hellas, Heraklion, Greece; and a Visiting Scholar with the Institute of Systems Research, University of Maryland at College Park, College Park, MD, USA. From 2012 to 2014, he was a Post-Doctoral Researcher with the Department of Telecommunications, CentraleSupélec, Gif-sur-Yvette, France. He is currently an Associate Professor with the Department of Computer and Information Science, Linköping University, Linköping, Sweden. His main research interests include the field of wireless communication networks with an emphasis on semantics-aware communications, energy harvesting networks, network-level cooperation, age of information, and stochastic geometry. He has served as the Symposium Co-Chair for the IEEE International Conference on Communications in 2022 and the IEEE Wireless Communications and Networking Conference in 2022. From 2013 to 2018, he was an Editor of the IEEE COMMUNICATIONS LETTERS. He was a Guest Editor of the IEEE INTERNET OF THINGS JOURNAL on "Age of Information and Data Semantics for Sensing, Communication and Control Co-Design in IoT." He is also an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING, the IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS, and the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, and an Expert Editor for invited papers of the IEEE COMMUNICATIONS LETTERS.



Marian Codreanu (Member, IEEE) received the M.Sc. degree from the University Politehnica of Bucharest, Romania, in 1998, and the Ph.D. degree from the University of Oulu, Finland, in 2007. His thesis was awarded as the Best Doctoral Thesis within the area of all technical sciences in Finland in 2007. In 2008, he was a Visiting Post-Doctoral Researcher with the University of Maryland, College Park, MD, USA. In 2013, the Academy of Finland awarded him a five years Academy Research Fellow position. In 2019, he received the Marie Skłodowska-Curie Individual Fellowship and joined Linköping University, where he is currently an Associate Professor. He published over 100 journals and conference papers in the areas of wireless communications and networking, statistical signal processing, mathematical optimization, and information theory. His current research interests include information freshness optimization, sparse signal processing, and machine learning for wireless networking.