# Minimizing the AoI in Multi-Source Two-Hop Systems under an Average Resource Constraint

Abolfazl Zakeri, Mohammad Moltafet, and Markus Leinonen
Centre for Wireless Communications – Radio Technologies
University of Oulu, Finland
e-mail: {abolfazl.zakeri, mohammad.moltafet, markus.leinonen}@oulu.fi

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

***Abstract*—We develop online scheduling policies to minimize the sum average age of information (AoI) subject to transmission capacity and long-run average resource constraints in a multi-source two-hop system, where independent sources randomly generate status update packets which are sent to the destination via a relay through error-prone links. A stochastic optimization problem is formulated and solved in known and unknown environments. For the known environment, an online near-optimal low-complexity policy is developed using the drift-plus-penalty method. For the unknown environment, a deep reinforcement learning policy is developed by employing the Lyapunov optimization theory and a dueling double deep Q-network. Simulation results show up to $136\%$ performance improvement of the proposed policy compared to a greedy-based baseline policy.**

***Index Terms*—Age of information (AoI), multi-source, two-hop, drift-plus-penalty, deep reinforcement learning.**

## I. Introduction

Timely delivery of status updates of a remotely monitored process to a destination is needed to support the emerging time-critical applications in the future Internet of things (IoT) in 5G and 6G wireless generations, e.g., industrial control, smart home systems, and drone control. *Age of information* (AoI) has been proposed to characterize the information freshness in status update systems [1]. The AoI is defined as the time elapsed since the latest received status update packet was generated [1], [2]. Nowadays, the AoI has attracted much interest in different areas, e.g., queuing systems [3]–[5], and scheduling and sampling problems [6]–[14]. The reader can refer to [15] for an extensive survey on the AoI.

We consider a *multi-source* two-hop status update system with *stochastic arrivals* and capacity limited *error-prone* (wireless) links. The sources independently generate different types of status update packets which arrive at a buffer-aided transmitter. The transmitter transmits the packets to a buffer-aided relay which further forwards them to the destination. The considered setup emerges in, e.g., vehicular networks, where status updates about various physical processes related to a vehicle are sent to a controller (e.g., a road side unit) to support vehicle safety applications; here, the communication from the vehicle to the controller requires the use of a relay, which could be another vehicle or an unmanned aerial vehicle (see [16] and examples therein). We provide online scheduling policies aiming at minimizing the sum average AoI (AAoI) subject to transmission capacity constraints and a constraint on the average number of all transmissions in the system. A stochastic optimization problem is formulated and solved in two different scenarios regarding the knowledge of system statistics: (i) known and (ii) unknown environments. For the known environment, a near-optimal *low-complexity drift-plus-penalty-based scheduling policy (DPP-SP)* is developed using drift-plus-penalty method [17]. For the unknown environment, we propose a deep reinforcement learning algorithm by employing the Lyapunov optimization theory and a dueling double deep Q-network (D3QN). Finally, simulation analysis are provided to examine the performance of the proposed scheduling policies; the results show up to $136\%$ performance improvement compared to a greedy-based baseline policy.

**Related Work:** Recently, the AoI in relaying systems has been studied in, e.g., [6], [9], [11], [18]–[25]. The work [6] studied the AoI minimization in a multi-source relaying system with the *generate-at-will* model (i.e., possibility of generating a new update at any time) and error-prone channels. In [9], the authors studied the AoI in a single-source energy harvesting relaying system with error-free channels and designed offline and online age-optimal policies. In [22], the authors considered a single-source relaying system under stochastic packet arrivals where the source communicates with the destination either through the direct link or via a relay. They proposed two different relaying protocols and derived the respective AAoI expressions. In summary, only a few works, such as [6], [9], [10], have incorporated a resource constraint (as we do in this paper) when analyzing the AoI in a relaying system, and most of the related works, e.g., [9], [18], [20], [22], consider single-source relaying systems.

Our work is an extension of work [7], where the authors provided scheduling policies for minimizing the AoI in a *one-hop buffer-free* network with stochastic arrivals and an *error-free* link. In contrast, in our two-hop network, all links are error-prone, and we further consider an average resource constraint. The most related work to our paper is [10], where the authors studied the AoI minimization problem in a *single-source* relaying system with the generate-at-will model under a resource constraint on the average number of forwarding transmissions at the relay. Different to [10], we consider a multi-source setup and, also, stochastic arrivals, which generalize the generate-at-will model adopted in [10]. Even though [10] also develops a low-complexity double threshold relaying

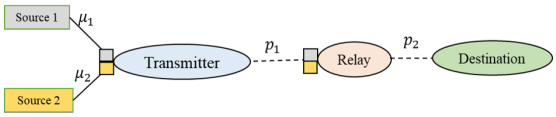



Fig. 1: A multi-source two-hop status update system in which different status updates arrive at random time slots at the transmitter, which then sends the packets to the destination via a buffer-aided relay over unreliable links.

policy, the thresholds need to be optimized numerically. In contrast, we provide: 1) an online near-optimal low-complexity scheduling policy, i.e., DPP-SP, and 2) a deep reinforcement learning policy that copes with unknown environment.

## II. System Model and Problem Formulation

### A. System Model

As depicted in Fig. 1, we consider a status update system consisting of two independent sources, a buffer-aided transmitter, a buffer-aided relay, and a destination. We consider two sources for simplicity of presentation. We assume that each status update is encapsulated in one packet and the buffer size in the transmitter and the relay is one packet per source. There is no direct communication link between the transmitter and the destination, and thus, the transmitter sends all status update packets to the destination via the relay. Each buffer stores the most recently arrived packet of a source to maintain the freshest information.

We consider a discrete-time system with unit time slots $t \in \{0, 1, 2, \ldots\}$. The sources, indexed by $i \in \{1, 2\}$, independently generate status update packets according to the Bernoulli distribution with parameter $\mu_i$. Let $u_i[t]$ be a binary indicator that shows whether a packet from source $i$ arrives at the transmitter at the beginning of slot $t$, i.e., $u_i[t] = 1$ indicates that a packet arrived; otherwise, $u_i[t] = 0$. Accordingly, $\Pr\{u_i[t] = 1\} = \mu_i$.

*Wireless Channels:* As the wireless channels fluctuate over time, reception of updates (both by the relay and the destination) are subject to errors. However, unsuccessfully received packets can be retransmitted; we assume that all retransmissions have the same reception success probability. Let $p_1$ and $p_2$ be the successful transmission probabilities of the transmitter-relay and relay-destination links, respectively. Also, let $\rho_1[t]$ be a binary indicator of a successful packet reception by the relay in slot $t$, i.e., $\rho_1[t] = 1$ indicates that the transmitted packet is successfully received by the relay; otherwise, $\rho_1[t] = 0$. Similarly, let $\rho_2[t]$ be a binary indicator of a successful packet reception by the destination in slot $t$, i.e., $\rho_2[t] = 1$ indicates that the transmitted packet is successfully received by the destination; otherwise, $\rho_2[t] = 0$. We have $\Pr\{\rho_1[t] = 1\} = p_1$ and $\Pr\{\rho_2[t] = 1\} = p_2$. We assume that instantaneous and error-free feedback is available for each link, and there is no interference between the links.

*Decision Variables:* We assume that at most one transmission per slot is possible over each link. Let $\alpha[t] \in \{0, 1, 2\}$ denote the (transmission) decision of the transmitter in slot $t$, where $\alpha[t] = i$, $i \in \{1, 2\}$, means that the transmitter sends the packet of source $i$ to the relay, and $\alpha[t] = 0$ means that the transmitter stays idle. Similarly, $\beta[t] \in \{0, 1, 2\}$ denotes the relay's decision in slot $t$, where $\beta[t] = i$, $i \in \{1, 2\}$, means that the relay forwards the packet of source $i$ to the destination, and $\beta[t] = 0$ means that the relay stays idle. We assume that there is a centralized controller performing the scheduling.

*Age of Information:* Let $\theta_i[t]$ denote the AoI of source $i$ at the transmitter in slot $t$. Also, let $\psi_i[t]$ denote the AoI of source $i$ at the relay and $\delta_i[t]$ denote the AoI of source $i$ at the destination in slot $t$. We make a common assumption (see e.g., [12], [13] and references therein) that AoI values are upper-bounded by a finite value $N$. This accounts for the fact that once the available information about the process of interest becomes excessively stale, further counting would be irrelevant. The evolution of the AoIs is given as

$$\theta_i[t+1] = \begin{cases} 0, & \text{if } \ u_i[t+1] = 1, \\ \min\big(\theta_i[t] + 1, N\big), & \text{otherwise}, \end{cases}$$

$$\psi_i[t+1] = \begin{cases} \min\big(\theta_i[t] + 1, N\big) & \text{if } \ \alpha[t] = i, \ \rho_1[t] = 1, \\ \min\big(\psi_i[t] + 1, N\big), & \text{otherwise}, \end{cases}$$

$$\delta_i[t+1] = \begin{cases} \min\big(\psi_i[t] + 1, N\big), & \text{if } \ \beta[t] = i, \ \rho_2[t] = 1, \\ \min\big(\delta_i[t] + 1, N\big), & \text{otherwise}. \end{cases}$$

### B. Problem Formulation

Let $\mathcal{D} = \{\alpha[t], \beta[t]\}_{t=1,2,3,\ldots}$ be a sequence of decision variables in the system. For a given $\mathcal{D}$, we denote the *sum average AoI at the destination* (S-AAoI) by $\bar{\delta}(\mathcal{D})$ and the *average number of total transmissions per slot* in the system by $\bar{K}(\mathcal{D})$, which are defined as

$$\bar{\delta}(\mathcal{D}) \triangleq \limsup_{T\to\infty} \ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{\delta_1[t] + \delta_2[t]\right\},$$

$$\bar{K}(\mathcal{D}) \triangleq \limsup_{T\to\infty} \ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{\mathbb{1}_{\{\alpha[t]\neq 0\}} + \mathbb{1}_{\{\beta[t]\neq 0\}}\right\},$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function which equals to 1 when the condition in $\{\cdot\}$ holds, and $\mathbb{E}\{\cdot\}$ is the expectation with respect to the system randomness (i.e., random wireless channels and packet arrival processes) and the decision variables $\alpha[t]$ and $\beta[t]$. The main aim of the paper is to solve the following stochastic optimization problem

$$\underset{\mathcal{D}}{\text{minimize}} \quad \bar{\delta}(\mathcal{D}) \tag{1a}$$

$$\text{subject to} \quad \bar{K}(\mathcal{D}) \leq \Gamma_{\max}, \tag{1b}$$

where the real value $\Gamma_{\max} \in (0, 2]$ is the maximum allowable average number of transmissions per slot in the system. Notice that the *Slater condition* [26, Eq. 9.32] clearly holds for problem (1), i.e., there exists some $\mathcal{D}$ for which $\bar{K}(\mathcal{D}) < \Gamma_{\max}$.

Problem (1) could be cast as a CMDP problem, and structural analysis of optimal policies can be conducted [8, Sec. III]. However, such approach suffers from the curse of dimensionality problem. Our focus is to provide a low-complexity scheduling policy to solve the main problem (1).

## III. Online Low-Complexity Scheduling Policy

In this section, we devise drift-plus-penalty-based scheduling policy, DPP-SP, inspired by the drift-plus-penalty method

[17], to solve the main problem (1). The proposed DPP-SP is a *heuristic*[1] policy that has low complexity and, as empirically shown in Section V, obtains near-optimal performance.

According to the drift-plus-penalty method [17], the time average constraint (1b) is enforced by transforming it into queue stability constraint. Accordingly, a virtual queue is associated for constraint (1b) in such a way that the stability of the virtual queue implies satisfaction of the constraint. Let $H[t]$ denote the virtual queue, with $H[0] = 0$, associated with constraint (1b) in slot $t$ which evolves as

$$H[t+1] = \max\{H[t] - \Gamma_{\max} + D(\mathbf{a}[t]),\ 0\}, \tag{2}$$

where $D(\mathbf{a}[t]) = \mathbb{1}_{\{\alpha[t]\neq 0\}} + \mathbb{1}_{\{\beta[t]\neq 0\}}$. By [17, Ch. 2], the time average constraint (1b) is satisfied if the virtual queue is *strongly stable*, i.e., $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{H[t]\} < +\infty$. Next, we define the Lyapunov function and its drift which are used to define the virtual queue stability condition.

We define a quadratic Lyapunov function as $L(H[t]) = \frac{1}{2}H^2[t]$ [17, Ch. 3]. The Lyapunov function indicates the size of the virtual queue: if the Lyapunov function is small, the virtual queue is small, and if the Lyapunov function is large, the virtual queue is large. By minimizing the expected change of the Lyapunov function from one slot to the next, the virtual queue can be stabilized [17, Ch. 4]. Let $\mathcal{Z}[t] = \{\mathbf{s}[t], H[t]\}$ denote the system state in slot $t$, where $\mathbf{s}[t] \triangleq (\theta_1[t], x_1[t], y_1[t], \theta_2[t], x_2[t], y_2[t])$ in which $x_i[t] \triangleq \psi_i[t] - \theta_i[t]$ and $y_i[t] \triangleq \delta_i[t] - \psi_i[t]$ are the *relative AoIs* at the relay and the destination in slot $t$, respectively. The one-slot *conditional Lyapunov drift*, $\Delta[t]$, is defined as the expected change in the Lyapunov function over one slot given the current system state $\mathcal{Z}[t]$, thus given as

$$\Delta[t] = \mathbb{E}\left\{L(H[t+1]) - L(H[t]) \;\middle|\; \mathcal{Z}[t]\right\}, \tag{3}$$

where the expectation is with respect to the (possibly random) decisions made in reaction to the current system state.

Applying the drift-plus-penalty method to main problem (1), we seek for a control policy that minimizes an upper bound on the following drift-plus-penalty function, $\varphi[t]$, at every slot:

$$\begin{aligned}&\varphi[t] = \Delta[t] + V\textstyle\sum_i \mathbb{E}\{\delta_i[t+1] + \psi_i[t+1]\big|\mathcal{Z}[t]\} = \\ &\Delta[t] + V\textstyle\sum_i \mathbb{E}\left\{(2\theta_i[t+1] + 2x_i[t+1] + y_i[t+1])\,\big|\mathcal{Z}[t]\right\},\end{aligned} \tag{4}$$

where the expectation is with respect to the channel randomness (i.e., $\rho_1[t]$ and $\rho_2[t]$) and (possibly random) decisions made in reaction to the current system state; parameter $V \geq 0$ adjusts a trade-off between the size of the virtual queue and the objective function.

To obtain the upper bound of the drift-plus-penalty function, we derive an upper bound for the drift term $\Delta[t]$, given by the following proposition.

**Proposition 1.** *The upper bound for the conditional Lyapunov drift in* (3) *is given by*

$$\Delta[t] \leq B + H[t]\big(\mathbb{E}\{D(\mathbf{a}[t]) \;\big|\; \mathcal{Z}[t]\} - \Gamma_{\max}\big), \tag{5}$$

*where* $B = 1/2\Gamma_{\max}^2 + 2$.

*Proof.* See [8, Appendix C]. □

Let us express the evolution of the AoIs and the relative AoIs by the following compact formulas

$$\begin{aligned}&\theta_i[t+1] = (1 - u_i[t+1])(\theta_i[t] + 1),\\ &x_i[t+1] = (1 - \rho_1[t]\mathbb{1}_{\{\alpha[t]=i\}})x_i[t] + u_i[t+1](\theta_i[t] + 1),\\ &y_i[t+1] = (1 - \rho_2[t]\mathbb{1}_{\{\beta[t]=i\}})y_i[t] + \rho_1[t]\mathbb{1}_{\{\alpha[t]=i\}}x_i[t].\end{aligned} \tag{6}$$

Using Proposition 1 and substituting (6) into (4), the upper bound for the drift-plus-penalty function $\varphi[t]$ is given as

$$\begin{aligned}\varphi[t] \leq\ & B + H[t](\mathbb{E}\{D(\mathbf{a}[t]) \;\big|\; \mathcal{Z}[t]\} - \Gamma_{\max})\\ &+V\textstyle\sum_i \big(\mathbb{E}\{(1 - \rho_2[t]\mathbb{1}_{\{\beta[t]=i\}})y_i[t]\\ &+(1 - \rho_1[t]\mathbb{1}_{\{\alpha[t]=i\}})x_i[t] + x_i[t] + 2\theta_i[t] + 2 \;\big|\; \mathcal{Z}[t]\}\big).\end{aligned} \tag{7}$$

Now, we turn to minimize the upper bound of the drift-penalty-function given in (7). To this end, we first compute the expectations with respect to the channel randomness, i.e., we have $\mathbb{E}\{\rho_2[t]\mathbb{1}_{\{\beta[t]=i\}} \mid \mathcal{Z}[t]\} = p_2\mathbb{E}\{\mathbb{1}_{\{\beta[t]=i\}} \mid \mathcal{Z}[t]\}$ and $\mathbb{E}\{\rho_1[t]\mathbb{1}_{\{\alpha[t]=i\}} \mid \mathcal{Z}[t]\} = p_1\mathbb{E}\{\mathbb{1}_{\{\alpha[t]=i\}} \mid \mathcal{Z}[t]\}$. Then, after removing the terms in (7) that are independent of the decision variables, we need to minimize the following expression:

$$\begin{aligned}&H[t]\mathbb{E}\{\mathbb{1}_{\{\beta[t]\neq 0\}}\big|\mathcal{Z}[t]\} - Vp_2\textstyle\sum_i \mathbb{E}\{\mathbb{1}_{\{\beta[t]=i\}}\big|\mathcal{Z}[t]\}y_i[t]\\ &+H[t]\mathbb{E}\{\mathbb{1}_{\{\alpha[t]\neq 0\}}\big|\mathcal{Z}[t]\} - Vp_1\textstyle\sum_i \mathbb{E}\{\mathbb{1}_{\{\alpha[t]=i\}}\big|\mathcal{Z}[t]\}x_i[t],\end{aligned} \tag{8}$$

where the expectation is with respect to the (possibly random) decisions.

To minimize the expression in (8), we follow the approach of opportunistically minimizing a (conditional) expectation [17, p. 13], i.e., the expression in (8) is minimized by the algorithm that observes the current system state $\mathcal{Z}[t]$ and chooses $\alpha[t]$ and $\beta[t]$ to minimize

$$\begin{aligned}&H[t]\mathbb{1}_{\{\alpha[t]\neq 0\}} - Vp_1\textstyle\sum_i \mathbb{1}_{\{\alpha[t]=i\}}x_i[t]\\ &+H[t]\mathbb{1}_{\{\beta[t]\neq 0\}} - Vp_2\textstyle\sum_i \mathbb{1}_{\{\beta[t]=i\}}y_i[t].\end{aligned} \tag{9}$$

The expression in (9) is separable in $\alpha[t]$ and $\beta[t]$, thus we obtain $\alpha[t]$ and $\beta[t]$ by solving the following problems

$$\underset{\alpha[t]\in\{0,1,2\}}{\text{minimize}}\quad H[t]\mathbb{1}_{\{\alpha[t]\neq 0\}} - Vp_1\textstyle\sum_i \mathbb{1}_{\{\alpha[t]=i\}}x_i[t], \tag{10}$$

$$\underset{\beta[t]\in\{0,1,2\}}{\text{minimize}}\quad H[t]\mathbb{1}_{\{\beta[t]\neq 0\}} - Vp_2\textstyle\sum_i \mathbb{1}_{\{\beta[t]=i\}}y_i[t]. \tag{11}$$

It can be inferred from problem (10) that if $H[t] \geq \max_i\{Vp_1x_i[t]\}$, then the optimal action is $\alpha[t] = 0$; otherwise, the optimal action is $\alpha[t] = \arg\max_i\{Vp_1x_i[t]\}$. Problem (11) has the similar solution with respect to $\beta[t]$.

In summary, the proposed DPP-SP works as follows: at each slot $t$, the controller observes $\mathcal{Z}[t]$ and determines the transmission decision variables according to the transmission decision rules given by (12), shown at the top of the next page. As seen in (12), DPP-SP performs only two simple operations to determine the actions at each slot, and thus,

[1]Even though the drift-plus-penalty method introduced in [17, Ch. 4] is guaranteed to give an asymptotically optimal policy, there is no guarantee on the optimally of DPP-SP because of having non-i.i.d. (over slots) objective function, i.e., the sum AoI at the destination.

$$\text{Structure of DPP-SP: } \begin{cases} \text{If } \max_i\{Vp_1x_i[t]\} \geq H[t], \text{ then } \alpha[t] = \arg\max_i\{Vp_1x_i[t]\}; \text{ otherwise, } \alpha[t] = 0, \\ \text{If } \max_i\{Vp_2y_i[t]\} \geq H[t], \text{ then } \beta[t] = \arg\max_i\{Vp_2y_i[t]\}; \text{ otherwise, } \beta[t] = 0. \end{cases} \tag{12}$$

DPP-SP has low complexity and can support systems with large numbers of sources.

What remains is to show that DPP-SP, operating according to (12), satisfies constraint (1b); this is shown in the following theorem.

**Theorem 1.** *For any finite $V$ and $N$, the virtual queue under DPP-SP that operates according to* (12) *is strongly stable, implying that DPP-SP satisfies constraint* (1b)*.*

*Proof.* See [8, Appendix D]. □

## IV. A Deep Learning Algorithm

In this section, we develop a deep (reinforcement) learning algorithm to solve the main problem (1) in an unknown environment, i.e., when the packet arrival rates and the successful transmission probabilities of the (wireless) links are not available for the controller. Due to the time average constraint (1b), we first use the Lyapunov optimization theory to convert the main problem (1) into an MDP problem which is then solved by a model-free deep learning algorithm, namely, D3QN (i.e., dueling double deep Q-network) [27]. While there is no guarantee that the proposed deep learning algorithm provides an optimal policy to the main problem (1), its advantage is that it can cope with unknown environments. We note that to implement the proposed learning algorithm, we do not need to bound the AoI values.

We define the expected time average reward function, obtained by policy $\pi$, as

$$R(\pi) \triangleq \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}\left\{r[t]\right\}, \tag{13}$$

where $r[t] = -\Big(L(H[t+1]) - L(H[t]) + V\sum_i \delta_i[t+1]\Big)$ is the *immediate reward function*, and $L(H[t]) = \frac{1}{2}H^2[t]$ is the quadratic Lyapunov function with virtual queue $H[t]$ given by (2). Now, we want to solve the following problem

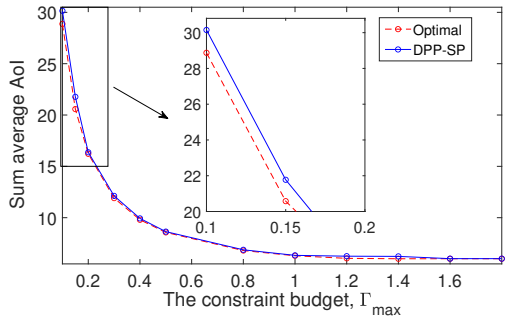
$$\underset{\pi}{\text{maximize}} \quad R(\pi). \tag{14}$$

Problem (14) can be formulated as an MDP problem, where $r[t]$ is the immediate reward, the state is $\mathcal{Z}[t] = \{\mathbf{s}[t], H[t]\}$, and the action is $\mathbf{a}[t] = (\alpha[t], \beta[t])$. To solve the MDP problem, we apply D3QN. Implementation details are presented in the next section. Finally, because the Slater condition holds and the AoI values are bounded by a finite $N$, it can be shown that the following theorem follows.

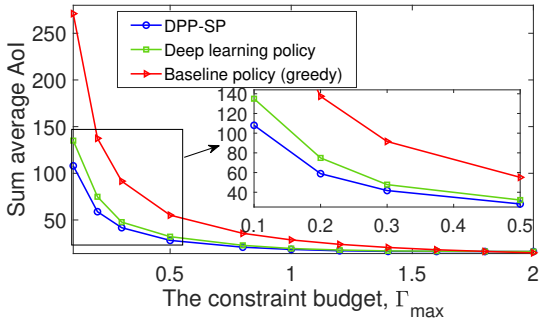
**Theorem 2.** *For any finite $V$ and $N$, the deep learning policy, i.e., D3QN, which solves (not necessarily optimally) problem* (14)*, makes the virtual queue be strongly stable, implying that it satisfies the time average constraint* (1b)*.*

## V. Numerical Results

In this section, we numerically evaluate the S-AAoI (i.e., sum average AoI at the destination) of the proposed policies. For the deep learning policy, we consider a fully-connected deep neural network consisting of an input layer ($|\mathcal{Z}[t]| = 6+1 = 7$ neurons), 2 hidden layers with 512 and 256 neurons and *ReLU* activation function, and an output layer ($|\mathcal{A}| = 9$ neurons). The number of steps per episode is 600, the discount factor is 0.99, the mini-batch size is 64, the learning-rate is 0.0001, and the optimizer is *RMSProp*. The parameter $V$ is set to 100. The results are averaged over 100,000 time slots. The system parameters, i.e., the arrival rates $\boldsymbol{\mu} = (\mu_1, \mu_2)$, the channel reliabilities $\mathbf{p} = (p_1, p_2)$, and the constraint budget $\Gamma_{\max}$ are specified in the figure captions.




(a) The S-AAoI v.s. constraint budget in a single-source setup for $\mathbf{p} = (0.3, 0.5)$ and $\mu_1 = 0.5$




(b) The S-AAoI v.s. constraint budget for $\mathbf{p} = (0.3, 0.5)$ and $\boldsymbol{\mu} = (0.4, 0.6)$

Fig. 2: The S-AAoI of the proposed and the baseline policies.

For benchmark, we consider a greedy "baseline policy", which determines the transmission decision variables at each slot $t$ according to the following rule: *If* $\bar{D}_t \leq \Gamma_{\max}$, *then* $\alpha[t] = \arg\max_i x_i[t]$ *and* $\beta[t] = \arg\max_i y_i[t]$*; otherwise,* $\alpha[t] = 0$ *and* $\beta[t] = 0$, where $\bar{D}_t$ denotes the average number of transmissions until slot $t$. This policy satisfies the time average constraint (1b). It is worth noting that the baseline policy and DPP-SP have similar computational complexity.

For a single-source setup, Fig. 2(a) compares the performance of DPP-SP (with $N = 90$) against an optimal policy, which is obtained by solving the linear program [26, Ch. 4] associated with the CMDP problem of the main problem (1) (see [8, Sec. III]). One source is considered for the computational tractability of the linear programming. The figure reveals that DPP-SP has near-optimal performance because it well

coincides with an optimal policy.

Fig. 2(b) depicts the S-AAoI of the proposed policies and the baseline policy as a function of the constraint budget $\Gamma_{\max}$ for two sources and unbounded AoIs. The figure shows that the deep learning policy obtains near-optimal performance when the constraint budget becomes sufficiently large, e.g., $\Gamma_{\max} \geq 0.8$; however, it is more complex (mainly in terms of training) than DPP-SP. Besides, the S-AAoI performance gap between the baseline policy and the proposed policies is extremely large when the constraint budget is small; this is because in such cases, performing good actions in each slot becomes more critical due to having a high limitation on the average number of allowed transmissions. The figure shows that the proposed policies achieve up to almost $136\%$ improvement in the S-AAoI performance compared to the baseline policy. Finally, we observe that, as the constraint budget increases, the S-AAoI values decrease; however, from a certain point onward, increasing the constraint budget does not considerably decrease the S-AAoI.

## VI. Conclusion

We developed transmission scheduling policies in a multi-source two-hop system with stochastic arrivals and error-prone channels subject to the transmission capacity and average number of transmissions constraints. We formulated a stochastic optimization problem and solved it in known and unknown environments. For the known environment, we devised the online near-optimal low-complexity DPP-SP, representing an efficient online scheduler for systems with large number of sources. For the unknown environment, we devised a deep learning policy combining the Lyapunov optimization theory and D3QN. The simulation results showed the effectiveness of the proposed policies, obtaining up to $136\%$ improvement in the S-AAoI performance compared to a greedy-based baseline policy. Thus, an age-optimal scheduler design is crucial for resource-constrained two-hop status update systems, where greedy-based scheduling is inefficient.

## VII. Acknowledgments

This research has been financially supported by the Infotech Oulu, the Academy of Finland (grant 323698), and Academy of Finland 6G Flagship program (grant 346208). The work of M. Leinonen has also been financially supported in part by the Academy of Finland (grant 340171).

## References

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

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

[3] M. Moltafet, M. Leinonen, and M. Codreanu, "On the age of information in multi-source queueing models," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5003–5017, Aug. 2020.

[4] E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *in Proc. IEEE Inter. Symp. on Inf. Theory (ISIT)*, pp. 131–135, Aachen, Germany, Jun. 2017.

[5] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1897–1910, Apr. 2016.

[6] J. Song, D. Gunduz, and W. Choi, "Optimal scheduling policy for minimizing age of information with a relay," *arXiv, preprint arXiv:2009.02716.*, pp. 1–1, Sep. 2020.

[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] A. Zakeri, M. Moltafet, M. Leinonen, and M. Codreanu, "Dynamic scheduling for minimizing AoI in resource-constrained multi-source relaying systems with stochastic arrivals," *arXiv preprint arXiv:2203.05656*, Mar. 2022.

[9] A. Arafa and S. Ulukus, "Timely updates in energy harvesting two-hop networks: Offline and online policies," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4017–4030, Aug. 2019.

[10] Y. Gu, Q. Wang, H. Chen, Y. Li, and B. Vucetic, "Optimizing information freshness in two-hop status update systems under a resource constraint," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1380–1392, May, 2021.

[11] J. Zhang and Y. Xu, "Age of information in relay-assisted status updating systems," *arXiv preprint arXiv:2107.01833*, Jul. 2021.

[12] 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.

[13] B. Zhou and W. Saad, "Minimum age of information in the internet of things with non-uniform status packet sizes," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1933–1947, Mar. 2020.

[14] J. Lou, X. Yuan, S. Kompella, and N.-F. Tzeng, "Boosting or hindering: AoI and throughput interrelation in routing-aware multi-hop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 29, no. 3, pp. 1008–1021, Jun. 2021.

[15] R. D. Yates, Y. Sun, D. Richard Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May, 2021.

[16] Z. Su, Y. Hui, T. H. Luan, Q. Liu, and R. Xing, *The Next Generation Vehicular Networks, Modeling, Algorithm and Applications*. Springer, 2020.

[17] M. J. Neely, *Stochastic network optimization with application to communication and queueing systems*. Synth. Lectures Commun. Netw., vol. 3, no. 1, pp. 1–211, Jan. 2010.

[18] M. Moradian and A. Dadlani, "Age of information in scheduled wireless relay networks," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, pp. 1–6, Seoul, Korea (South), Mar. 2020.

[19] B. Li, H. Chen, N. Pappas, and Y. Li, "Optimizing information freshness in two-way relay networks," in *Proc. IEEE/CIC Int. Conf. on Commun. in China (ICCC)*, pp. 893–898, Chongqing, China, Aug. 2020.

[20] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Age-optimal information updates in multihop networks," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 576–580, Aachen, Germany, Jun. 2017.

[21] J. Feng, H. Pan, T.-T. Chan, and J. Liang, "Timely status update: Should ARQ be used in two-hop networks?," *arXiv preprint arXiv:2201.10253*, Jan. 2022.

[22] B. Li, Q. Wang, H. Chen, Y. Zhou, and Y. Li, "Optimizing information freshness for cooperative IoT systems with stochastic arrivals," *IEEE Internet Things J*, vol. 8, no. 19, pp. 14485–14500, Oct. 2021.

[23] J. P. Champati, H. Al-Zubaidy, and J. Gross, "Statistical guarantee optimization for AoI in single-hop and two-hop FCFS systems with periodic arrivals," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 365–381, Jun. 2021.

[24] J. Lou, X. Yuan, S. Kompella, and N.-F. Tzeng, "AoI and throughput tradeoffs in routing-aware multi-hop wireless networks," in *in Proc. IEEE Con. on Comp. Commun.*, pp. 476–485, Toronto, ON, Canada, Jul. 2020.

[25] D. Zheng, Y. Yang, L. Wei, and B. Jiao, "Decode-and-forward short-packet relaying in the internet of things: Timely status updates," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8423–8437, Dec. 2021.

[26] E. Altman, *Constrained Markov Decision Processes*. volume 7. CRC Press, 1999.

[27] L. Xie, S. Wang, A. Markham, and A. Trigoni, "Towards monocular vision based obstacle avoidance through deep reinforcement learning," *ArXiv*, vol. abs/1706.09829, Jun., 2017.