Jointly Minimizing AoI Penalty and Network Cost Among Coexisting Source-Destination Pairs

Cho-Hsin Tsai and Chih-Chun Wang
School of ECE, Purdue University; Email: {chtsai, chihw}@purdue.edu

***Abstract*—One main objective of ultra-low-latency communications is to minimize the *data staleness* at the receivers, recently characterized by a metric called Age-of-Information (AoI). While the question of *when to send the next update packet* has been the central subject of AoI minimization, each update packet also incurs the *cost of transmission* that needs to be jointly considered in a practical design. With the exponential growth of interconnected devices and the increasing risk of excessive resource consumption in mind, this work derives an optimal joint cost-and-AoI minimization solution for multiple coexisting source-destination (S-D) pairs. The results admit a new *AoI-market-price*-based interpretation and are applicable to the setting of (a) general heterogeneous AoI penalty functions and Markov delay distributions for each S-D pair, and (b) a general network cost function of aggregate throughput of all S-D pairs. Extensive simulation is used to demonstrate the superior performance of the proposed scheme.**

I. Introduction

The increasing demand for real-time communications, including VR/AR systems, remote surgeries, and autonomous driving services [1]–[3], prompts a back-to-basics approach for next-generation low-latency network designs [4]. Recently, a new metric called Age-of-Information (AoI) was introduced to rigorously quantify *data staleness* [5]. The corresponding AoI minimization problems have since been studied for various settings [6]–[11]. In particular, [12]–[14] considered a single source-destination (S-D) pair, where each data packet sent from the source experiences random delay. Once it is delivered to the destination, an instantaneous ACK packet will inform the source, and the ACK will be used to decide when to send the next packet. [12], [13] characterized the optimal transmission schedules that minimize the average AoI. [14] extends the results to arbitrary AoI penalty functions. Recently, [15] generalized the 1-way-delay setting in [12]–[14] to the 2-way-delay setting, where the ACK also experiences random delay, and proposed a distribution-oblivious online algorithm that provably converges to the optimum.

Meanwhile, as the exponential growth of connected devices brings more convenience to the society, such as gigantic number of devices also poses an increasing risk of excessive resource consumption [16]. Existing results [13], [14] assume a *maximum average sampling rate constraint* $R_{\max}$, which essentially solve[1] the following joint cost-&-AoI minimization problem

$$\text{minimize} \quad \mathsf{avg.aoi.penalty} + \mathsf{loss}(\mathsf{sampling.rate}) \quad (1)$$

where the network cost function $\mathsf{loss}(x) = 0$ if $x \le R_{\max}$ and $\mathsf{loss}(x) = \infty$ if $x > R_{\max}$. This work strengthens the results in [13], [14] with the following contributions.

(i) Instead of a specialized cost function $\mathsf{loss}(\cdot)$, we solve (1) for arbitrary continuously differentiable, convex, non-decreasing $\mathsf{loss}(\cdot)$. The results characterize the complete trade-off between cost and AoI under a more general throughput-to-cost and AoI-penalty structure, and hence significantly broaden the applicability of [13], [14].

(ii) We generalize the single S-D pair results [13], [14] for multiple coexisting S-D pairs. Specifically, we characterize how coexisting S-D pairs can optimally and collectively balance their individual AoI minimization goals via a shared cost function of the *aggregate network throughput*. See (7) for the rigorous definition. The solution takes into account different *timeliness requirements* and different *transmission costs* of each S-D pair by allowing for heterogeneous AoI penalty functions and general cost function $\mathsf{loss}(\cdot)$. This network-wide joint cost-&-AoI minimization will greatly benefit future 5G network designs, which aim to support a million devices in a square kilometer [17] of widely-ranging throughput-cost and AoI targets.

(iii) Analytically, the solution admits a new *AoI-market-price*-based interpretation, and can thus be viewed as a new AoI-centric network utility maximization (NUM) framework.

(iv) Simulation results show that our scheme successfully curbs the excessive resource consumption of the existing cost-oblivious AoI-optimal policy and optimally balances all S-D pairs with 24–56% savings compared to the state of the art.

A. Existing Joint Cost-&-AoI Minimization Results

This work focuses on the cost-&-AoI minimization for the queue-based setting with random service time and random ACK delay, see [13]–[15]. Existing cost-&-AoI minimization results are based on various significantly different network scenarios [13], [14], [18]–[25]. For instance, both [18] and [19] considered AoI minimization with either the *average power* constraint or with the *sampling cost* consideration. Both considered *block fading channels*, in which whether each packet transmission succeeds or fails will be fed back to the source instantaneously at the end of the time slot. There is no concept of *random delayed delivery* that is central in our setting. [24], [26], [27] focused on *energy harvesting sources*,

This work was supported in parts by NSF under Grant CCF-1422997, CCF-1618475, CCF-1816013 and CCF-2008527.

[1] While the results of [13], [14] are presented as having a maximum sampling rate constraint, the approaches of [13], [14] can also be used to solve the cost-&-AoI minimization problem when $\mathsf{loss}(x)$ is affine.

a scenario that is very different from our simple but highly relevant setting of multiple S-D pairs with delayed delivery. In addition, most existing works considered AoI with an affine throughput-to-cost function [13], [14], [20]–[23], [25], another distinction from the general $\mathsf{loss}(\cdot)$ in this work.

II. Model and Formulation

A. System Model With K Source-Destination Pairs

We consider a network of K coexisting S-D pairs, each of which is composed of a source, a destination, a source-to-destination (s2d) channel and a destination-to-source (d2s) channel as shown in Fig. 1. The *ACK-based generate-at-will* model [12]–[15] is considered. To be specific, after the source transmits a packet, the packet will experience some delay before arriving at the destination. Once delivered, the destination immediately generates an ACK-packet and sends it back to the source, which again may experience some delay. After the ACK of the previous packet arrives, the source can wait for an arbitrary amount of time. After the carefully chosen waiting time, the source generates a new status update packet and transmits it. The process then repeats itself. The detailed system evolution is described below (also see [13]–[15] for the description for the special case of $K = 1$).

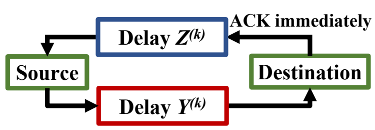

Fig. 1: An S-D pair with two-way delay.

Time sequences: For the k-th S-D pair, the system consists of three discrete-time real-valued non-negative random processes $X_i^{(k)}$, $Y_i^{(k)}$, and $Z_i^{(k)}$, for all $i \geq 0$. $X_i^{(k)}$ is the waiting time of the source between receiving the $(i-1)$-th ACK and generating/transmitting the i-th update packet;[2] $Y_i^{(k)}$ (resp. $Z_i^{(k)}$) is the random delay for the i-th use of the s2d (resp. d2s) channel.

For each S-D pair, $S_i^{(k)}$ denotes the time instant when the i-th packet is generated/transmitted. The i-th packet arrives at the destination at time $D_i^{(k)}$, and the i-th ACK packet is received by the source at time $A_i^{(k)}$. The values of $(S_i^{(k)}, D_i^{(k)}, A_i^{(k)})$ refer to the absolute time instants, while the values of $(X_i^{(k)}, Y_i^{(k)}, Z_i^{(k)})$ represent the lengths of the intervals. They are related by the following equations: Initialize $A_0^{(k)} = X_0^{(k)} = Y_0^{(k)} = Z_0^{(k)} = 0$. For all $i \geq 1$, we have $S_i^{(k)} = A_{i-1}^{(k)} + X_i^{(k)}$, $D_i^{(k)} = S_i^{(k)} + Y_i^{(k)}$, and $A_i^{(k)} = D_i^{(k)} + Z_i^{(k)}$. We call the time interval $[A_{i-1}^{(k)}, A_i^{(k)})$ the *i-th round*, consisting of the i-th waiting time $X_i^{(k)}$ at the source, the i-th s2d delay $Y_i^{(k)}$ and d2s delay $Z_i^{(k)}$. See Fig. 2.

[2]The assumption that $X_i^{(k)} \geq 0$ prevents the source from transmission before receiving the ACK, which reflects the principle in the stop-and-wait ARQ mechanism [28]. One may design an even better algorithm that transmits anticipatively before the ACK is delivered, which, however, is beyond the scope of this paper.

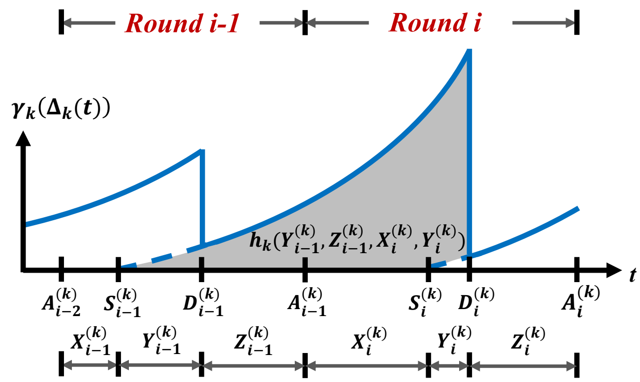

Fig. 2: Evolution of the AoI penalty function $\gamma_k(\Delta_k(t))$.

Age-of-Information and its penalty function: Following [5], we define the Age-of-Information $\Delta_k(t)$ at time t by

$$\Delta_k(t) \triangleq t - \max\{S_i^{(k)} : i \text{ satisfies } D_i^{(k)} \leq t\}. \tag{2}$$

The *AoI penalty function* $\gamma_k(\Delta_k(t))$ represents the level of data staleness. Three popular choices are: (i) linear $\gamma_{k,\text{lin}}(\Delta) = w_k \cdot \Delta$ [29]; (ii) quadratic $\gamma_{k,\text{qrd}}(\Delta) = w_k \cdot \Delta^2$ [15]; and (iii) exponential $\gamma_{k,\text{exp}}(\Delta) = e^{w_k \Delta} - 1$ [14]. In all the three choices, $w_k \geq 0$ are tunable parameters. An S-D pair carrying time-sensitive traffic may use an exponential penalty function and/or use a larger weight w_k, while less urgent traffic may use a linear penalty and/or with a smaller w_k. Our results hold for any heterogeneous choices of $\gamma_k(\cdot)$, not limited to the above three. See Fig. 2 for the evolution of $\gamma_k(\Delta_k(t))$.

Overall objective: Define the k-th average throughput by

$$R_k(T) = \frac{1}{T}\mathbb{E}\{\max\{i : A_i^{(k)} \leq T\}\}. \tag{3}$$

We use a single $\mathsf{loss}(\cdot)$ function to represent the cost for the network *to carry the traffic of all K pairs.* We aim to minimize

$$\limsup_{T\to\infty} \left(\sum_{k=1}^{K} \frac{1}{T}\int_0^T \gamma_k(\Delta_k(t))dt\right) + \mathsf{loss}\left(\sum_{k=1}^{K} c_k R_k(T)\right) \tag{4}$$

where the constants $c_k > 0$ describe the (relative) amount of resource consumption for carrying the underlying traffic, e.g., a large c_k means it is more costly to carry the k-th S-D pair. Note that the AoI penalty $\gamma_k(\cdot)$ can also be individually weighted for each pair. We do not explicitly specify their weighting coefficients herein since they can be completely absorbed when choosing $\gamma_k(\cdot)$ arbitrarily and heterogeneously.

Technical assumptions (i)–(v): (i) $\mathsf{loss}(\cdot) : [0, \infty) \to (-\infty, \infty)$ is a continuously differentiable, non-decreasing and convex function; for each S-D pair, we assume (ii) $Y^{(k)}$ and $Z^{(k)}$ are of *bounded support*; (iii) $(Y_i^{(k)}, Z_i^{(k)})$ can be of arbitrary joint distribution $\mathbb{P}_{Y^{(k)}Z^{(k)}}$ but the vector random process $\{(Y_i^{(k)}, Z_i^{(k)}) : i \geq 1\}$ is ergodic, stationary and Markov; (iv) $\mathbb{E}\{Y_i^{(k)}\} + \mathbb{E}\{Z_i^{(k)}\} > 0$; (v) $\gamma_k(\cdot) : [0, \infty] \to [0, \infty]$ is a continuously differentiable and strictly increasing function satisfying $\gamma_k(0) = 0$ and $\gamma_k(\infty) = \infty$.

Our model implicitly assumes the network can support all the K users. However, we allow for any arbitrary convex loss function $\mathsf{loss}(\cdot)$ to implicitly deal with the inability of supporting too many users, e.g., if we set $\mathsf{loss}(\cdot)$ to be infinity when the number of users $> K_{\max}$, then the scheme would automatically limit the number of users. The loss function $\mathsf{loss}(\cdot)$ is a function of the "average" throughput, not the "instantaneous" throughput. Therefore, it is an instance of "soft constraints" (see similar rate constraints in [13], [14] that can also be viewed as instances of soft constraints).

B. From the Long-Term Average to a Single-Round Analysis

We first define two functions for the k-th S-D pair:

$$h_k(y', z', x, y) \triangleq \int_0^{y'+z'+x+y} \gamma_k(t)dt - \int_0^{y} \gamma_k(t)dt \tag{5}$$

$$g_k(y', z', x) \triangleq \mathbb{E}\left\{h_k(y', z', x, Y_i^{(k)})|Y_{i-1}^{(k)} = y', Z_{i-1}^{(k)} = z'\right\} \tag{6}$$

where $g_k(y', z', x)$ is the conditional expectation of $h_k(y', z', x, Y_i^{(k)})$ over $Y_i^{(k)}$. The intuition behind (5) is that the shaded area in Fig. 2 is computed by $h_k(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}, X_i^{(k)}, Y_i^{(k)})$. We observe that the overall area underneath $\gamma_k(\Delta_k(t))$ can be decomposed as a summation of smaller sub-areas with shapes similar to the shaded area $h_k(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}, X_i^{(k)}, Y_i^{(k)})$ in Fig. 2. Using this observation (also used in [13]–[15]), we can convert the original problem (4) to the following equivalent *single-round minimization problem*:

$$\mu^* \triangleq \inf_{X_i^{(1)},\dots,X_i^{(K)}} \sum_{k=1}^{K} \frac{\mathbb{E}\left\{g_k(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}, X_i^{(k)})\right\}}{\mathbb{E}\left\{Y_{i-1}^{(k)} + Z_{i-1}^{(k)} + X_i^{(k)}\right\}} + \mathsf{loss}\left(\sum_{k=1}^{K} \frac{c_k}{\mathbb{E}\left\{Y_{i-1}^{(k)} + Z_{i-1}^{(k)} + X_i^{(k)}\right\}}\right) \tag{7}$$

for which the i-th waiting time $X_i^{(k)}$ is a function of the delays of the previous round $(Y_{i-1}^{(k)}, Z_{i-1}^{(k)})$. Our goal is to design the set of K waiting time functions $\{X_i^{(1)}, ..., X_i^{(K)}\}$ that minimizes (7). The value of the round index i is irrelevant herein since whatever design that minimizes (7) can and will be repeatedly applied to all rounds $i \geq 1$. In the sequel, we focus exclusively on solving (7).

III. Main Results

Recall that the waiting time $X_i^{(k)}$ is a function of $(Y_{i-1}^{(k)}, Z_{i-1}^{(k)})$. We consider the following *hitting-time*-based $\phi_\beta^{(k)}$ with a tunable parameter $\beta \geq 0$:

$$X_i^{(k)}(\beta) = \phi_\beta^{(k)}(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}) \tag{8}$$

$$\triangleq \inf\left\{t > 0 : \frac{d}{dt} g_k(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}, t) > \beta\right\}. \tag{9}$$

Note that the waiting time function $\phi_\beta^{(k)}(\cdot,\cdot)$ is based on $g_k(\cdot)$, which implicitly depends on the given, likely heterogeneously chosen, k-th AoI penality function $\gamma_k(\cdot)$ and the k-th delay distribution $Y_i^{(k)}$, see (5) and (6). As a result, different k may have a different $\phi_\beta^{(k)}(\cdot,\cdot)$ even though they share the same form of (9). For each β, we further define

$$\mathsf{aoi}_k(\beta) \triangleq \mathbb{E}\left\{g_k(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}, \phi_\beta^{(k)}(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}))\right\} \tag{10}$$

$$T_k(\beta) \triangleq \mathbb{E}\left\{Y_{i-1}^{(k)} + Z_{i-1}^{(k)} + \phi_\beta^{(k)}(Y_{i-1}^{(k)}, Z_{i-1}^{(k)})\right\} \tag{11}$$

as the expected AoI penalty and time duration of the special scheme $\phi_\beta^{(k)}(\cdot,\cdot)$, where both expectations are taken over the random vector $(Y_{i-1}^{(k)}, Z_{i-1}^{(k)})$.

Lemma 1: Under technical assumptions (i)–(v), $\mathsf{aoi}_k(\beta)$ and $T_k(\beta)$ are continuous, non-decreasing and strictly positive functions of β.

By the definition of (9), if $\frac{d}{dt}g_k(\cdot,\cdot,t)|_{t=0}$ is lower bounded away from 0, then $\phi_\beta^{(k)}(\cdot,\cdot) = 0$ for all β that are sufficiently close-to-zero. Therefore, the values of $\mathsf{aoi}_k(\beta)$ and $T_k(\beta)$ in (10) and (11) do not change with respect to β when β is small. To formalize this observation, for every $1 \leq k \leq K$, we define

$$\beta_0^{(k)} \triangleq \max\{\beta : T_k(\beta) = T_k(0)\}. \tag{12}$$

We then have $T_k(\beta_1) = T_k(\beta_2)$ and $\mathsf{aoi}_k(\beta_1) = \mathsf{aoi}_k(\beta_2)$ for all $\beta_1, \beta_2 \in [0, \beta_0^{(k)}]$. Before proceeding, we introduce one more technical assumption.

Technical assumption (vi): For each k, we assume $\mathsf{aoi}_k(\beta)$ and $T_k(\beta)$ are continuously differentiable with respect to β if $\beta > \beta_0^{(k)}$.

In other words, we assume $\mathsf{aoi}_k(\beta)$ and $T_k(\beta)$ are well-behaved versus β while recognizing that $\mathsf{aoi}_k(\beta)$ (resp. $T_k(\beta)$) contains two pieces, one being a flat line segment for $\beta \in [0, \beta_0^{(k)}]$ and one being a non-decreasing curve for $\beta \in [\beta_0^{(k)}, \infty)$ (see Lemma 1). The left and right derivatives at the junction point $\beta = \beta_0^{(k)}$ are generally not equal. Hence the derivative continuity is only assumed for $\beta > \beta_0^{(k)}$.

Following (7), we define

$$\mu_{\text{HT}}(\beta_1, ..., \beta_K) \triangleq \sum_{k=1}^{K} \frac{\mathsf{aoi}_k(\beta_k)}{T_k(\beta_k)} + \mathsf{loss}\left(\sum_{k=1}^{K} \frac{c_k}{T_k(\beta_k)}\right) \tag{13}$$

as the objective value achieved *when all K pairs employ the hitting-time policy* (8) *and* (9) with thresholds being $(\beta_1, ..., \beta_K)$, respectively. Also, define

$$\mu_{\text{HT}}^* \triangleq \min_{\beta_1,\dots,\beta_K} \mu_{\text{HT}}(\beta_1, ..., \beta_K) \tag{14}$$

as the minimum value achieved if all K pairs use the hitting-time policy (after optimizing $(\beta_1, ..., \beta_K)$). It is clear that $\mu^* \leq \mu_{\text{HT}}^*$ since the latter is restricted to a special class of hitting-time-based policies.

Proposition 1: Under technical assumptions (i)–(vi), we have

$$\mu^* = \mu_{\text{HT}}^*. \tag{15}$$

A high-level proof sketch is as follows. Consider any fixed k. Suppose for an arbitrarily given scheme, the corresponding expected i-th round AoI-penalty is a_k and the expected i-th round time duration is t_k. We first prove that we can always find a $\beta_k \geq 0$ such that $\frac{\mathsf{aoi}_k(\beta_k)}{T_k(\beta_k)} \leq \frac{a_k}{t_k}$ and $\frac{1}{T_k(\beta_k)} \leq \frac{1}{t_k}$; that is, the hitting-time policy using such β_k "improves" the average AoI penalty and the average throughput simultaneously. Since the $\mathsf{loss}(\cdot)$ is non-decreasing, replacing the given scheme by our hitting-time policy will not hurt the overall objective value in (7), hence Proposition 1.

Proposition 1 explicitly proves that the best hitting-time policy attains the optimal objective function value of any possible designs. Therefore, in the following we can restrict our focus to the hitting-time policies without sacrificing optimality.

The following results describe how to compute the thresholds $(\beta_1, ..., \beta_K)$ that achieves the minimum $\mu^*_{\text{HT}} = \mu^*$.

Lemma 2: For each k,

$$\beta_k \cdot T_k(\beta_k) - \mathsf{aoi}_k(\beta_k) \tag{16}$$

is a continuous and strictly increasing function of β_k. Furthermore, its value is strictly negative when $\beta_k = 0$ and it approaches ∞ when $\beta_k \to \infty$.

By Lemma 2, $\forall x \geq 0$ the following equation

$$\frac{\beta_k \cdot T_k(\beta_k) - \mathsf{aoi}_k(\beta_k)}{c_k} = x \tag{17}$$

has a unique β_k solution in $(0, \infty)$, which we denote by $\beta_k(x)$ to emphasize its dependency on x. Define

$$m(r) \triangleq \frac{d}{dr}\mathsf{loss}(r) \tag{18}$$

as the slope of the cost function $\mathsf{loss}(\cdot)$. We then define

$$f(\beta_1, \cdots, \beta_K) \triangleq m\left(\sum_{k=1}^{K} \frac{c_k}{T_k(\beta_k)}\right). \tag{19}$$

Lemma 3: The following fixed-point equation

$$f(\beta_1(x), \cdots, \beta_k(x)) = x \tag{20}$$

has a unique root x^* in the interval $[0, f(\beta_1(0), \cdots, \beta_K(0))]$.

Proposition 2: Eq. (14) is attained by

$$\mu^*_{\text{HT}} = \mu_{\text{HT}}(\beta_1(x^*), ..., \beta_K(x^*)). \tag{21}$$

Note that all our results require technical assumptions (i) to (vi). For example, the uniqueness of x^* in Lemma 3 requires the convexity of $\mathsf{loss}(\cdot)$.

An AoI-market-price-based interpretation: The intuition behind Lemmas 2 and 3 and Proposition 2 is as follows. Given x, the solution $\beta_k(x)$ is the threshold parameter that leads to the hitting-time policy that *minimize the average AoI penalty of the k-th pair under the marginal cost x*. Consequently, x can be viewed as the price that the S-D pair has to pay for each packet transmission. The larger the price x, the larger the $\beta_k(x)$ (see Lemma 2 and (17)), the longer the expected duration of each packet transmission $T_k(\beta_k(x))$ (see Lemma 1), the less willing for each S-D pair to send a new packet.

As a result, the right-hand side of the fixed-point equation (20) is the "market price" of each packet transmission *each S-D pair is willing to pay* when operating under parameter $\beta_k(x)$. Note that $f(\beta_1, \cdots, \beta_K)$ in (18) and (19) is the *marginal* network cost if the network is to support all K pairs that adopt the hitting-time policy with parameters β_1 to β_K, respectively. The left-hand side of (20) is thus how much *the network would charge for each packet transmission* if all S-D pairs operate under $(\beta_1(x), \cdots, \beta_K(x))$. The optimum (equilibrium) is attained when the market price balances how much each S-D pair is willing to pay and how much the network has to charge for each packet transmission, hence the fixed-point equation.

By Propositions 1 and 2, we can find an optimal policy that achieves μ^* by the following steps.

Step 1: For each S-D pair, use the given AoI penalty function $\gamma_k(\cdot)$ and the delay distribution $(Y_i^{(k)}, Z_i^{(k)})$ to find the explicit expression of $h_k(\cdot)$, $g_k(\cdot)$, $\phi_\beta^{(k)}(\cdot)$ by (5), (6), and (9), respectively; **Step 2:** Use the waiting time function $\phi_\beta^{(k)}(\cdot)$ to derive the functions $\mathsf{aoi}_k(\beta)$ and $T_k(\beta)$ of β by (10) and (11) and the resulting function $\beta_k(x)$ of x by (17); **Step 3:** By Lemma 3, we use the bisection method [30] to find x^*, the root of the fixed-point equation (20), and then derive the optimal thresholds $(\beta_1(x^*), ..., \beta_K(x^*))$ in Proposition 2; **Step 4:** For the k-th S-D pair, if the $(i-1)$-th round delays are $(Y_{i-1}^{(k)}, Z_{i-1}^{(k)}) = (y', z')$, the source simply waits for $X_i^{(k)} = \phi_{\beta_k(x^*)}^{(k)}(y', z')$ amount of time, see (8) and (9), before generating/transmitting the next (i.e., the i-th) packet.

Steps 1 to 3 can be computed offline. Step 4 is a simple hitting-time-based policy using the computed parameter $\beta_k(x^*)$. It is worth pointing out that the K-pair jointly optimal solution can be computed very efficiently for large K, say $K = 10^3$, since the optimal K-dimensional vector $(\beta_1^*, \cdots, \beta_K^*)$ in (14) is found by solving a *1-dimensional fixed-point equation* (20) via bisection.

IV. Simulation Results

We compare our scheme to the following benchmarks:

(i) Zero-Wait (ZW) policy [31]: $X_i^{(k)} = 0$, $\forall k, i$. ZW is known to maximize the sum throughput [31].

(ii) AoI-Optimal policy [13]–[15]: this policy is also a hitting-time-based policy[3] but is oblivious to the network cost.

Due to the space limit, we only report the simulation results using log-normal delays, which are empirically reasonable channel models [32]. Similar results have been observed with other delay distributions.

A. The Case of $K = 1$

When $K = 1$, our results characterize the joint cost-&-AoI minimal solution of a single-pair setting with arbitrary cost function $\mathsf{loss}(\cdot)$, which has not been studied in [13], [14] that consider the maximum sampling rate constraint.

We consider a single S-D pair with delays Y and Z being log-normal random variables with $(\mu_Y, \sigma_Y^2) = (0.5, 0.25)$, $(\mu_Z, \sigma_Z^2) = (0.5, 0.5)$, and the correlation coefficient $\rho_{YZ} =$

[3]This is consistent with our findings since our results contain the cost-oblivious setting [13]–[15] as a special case once we hardwire $\mathsf{loss}(\cdot) = 0$.

0.66. The vector process $\{(Y_i, Z_i) : i \geq 1\}$ is i.i.d. and the AoI penalty function is quadratic $\gamma_1(\Delta) = 0.5 \cdot \Delta^2$. We consider an exponential $\text{loss}(r) = e^{\alpha r} - 1$ with $c_1 = 1$, see (7). The waiting time of the proposed policy can be computed using the four steps outlined in the end of Sec. III. For instance, when $\alpha = 16$, we have $x^* = 147.21$ and $\beta_1(x^*) = 39.37$ according to Steps 1 to 3. If the delays in the previous round are $(Y_{i-1}^{(1)} = 1, Z_{i-1}^{(1)} = 1)$, then the waiting time for this round is $X_i^{(1)}(\beta_1(x^*)) = 4.95$.

We run the three schemes ZW, AoI-Optimal, and Proposed scheme for different α values and Fig. 3 plots the resulting joint cost-&-AoI objective values. As expected, the Proposed always achieves the lowest objective value. For large $\alpha = 20$, it leads to substantial savings of 80% and 66% when compared to ZW and AoI-Optimal policies, respectively.

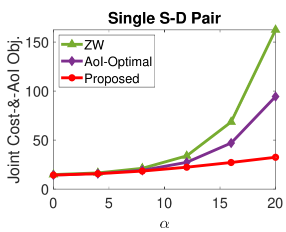

Fig. 3: Simulation results for a single S-D pair.

B. *The Case of $K = 5$ With Two Classes of Traffic*

We also examine the case when heterogeneous traffics are competing for the shared resources. In this experiment, we fix $\text{loss}(r) = e^{4r} - 1$ and consider five S-D pairs with two classes of traffic. We vary the composition of the two classes between 20/80 (one Class-1 S-D pair and four Class-2 S-D pairs) to 80/20 (four Class-1 pairs and one Class-2 pair).

All five S-D pairs have identical log-normal delays Y and Z as described in Sec. IV-A. However, each Class-1 pair has an AoI penalty function $\gamma_{\text{Class1}}(\Delta) = \Delta^2$, whereas each Class-2 pair has a smaller AoI penalty function $\gamma_{\text{Class2}}(\Delta) = 0.05 \cdot \Delta^2$; that is, Class-1 traffic is more "urgent." For simplicity, we do not impose individual throughput weighting and thus set $c_k = 1$ for all $k = 1$ to 5, see (7).

Compared to the cost-oblivious AoI-Optimal policy, the savings of the Proposed range from 56% for the 20/80 case to 24% for the 80/20 case, see Fig. 4a. For deeper analysis, we plot the single-pair average AoI penalty for both classes, see Figs. 4c and 4d. Since Class-1 has a larger AoI penalty function, its *average AoI penalty* is always higher (see the magnitude of of y-axis in Figs. 4c and 4d). Furthermore, since Class-1 is more "urgent," our scheme allocates higher throughput for Class-1 than Class-2 (see the magnitude of y-axis in Figs. 4e and 4f).

We also observe that when the percentage of Class-1 pairs increases, the network is carrying more "urgent" traffic. However, since all pairs share the same network resources, each Class-1 pair cannot expect to receive the same amount of bandwidth as before. As a result, the throughput of each Class-1 pair decreases (Fig. 4e) and its average AoI penalty increases (Fig. 4c). Each Class-2 pair also reduces its own throughput (Fig. 4f) and increases its own AoI penalty (Fig. 4d) to *make room for the newly added Class-1 pairs in the network.* Overall, the sum throughput of all 5 pairs increases when we have more Class-1 pairs (Fig. 4b). This is because the network recognizes that it is now carrying more urgent pairs and increases its sum throughput accordingly.[4]

Note that both ZW and AoI-Optimal are cost-oblivious and each S-D pair thus blindly decides its own transmission policy without considering the collective network resource consumption. That is why all their performance curves are flat lines that do not react to different Class-1 percentages. In contrast, the proposed scheme optimally balances the throughput and AoI-penalty while taking into account different compositions of the underlying traffic.

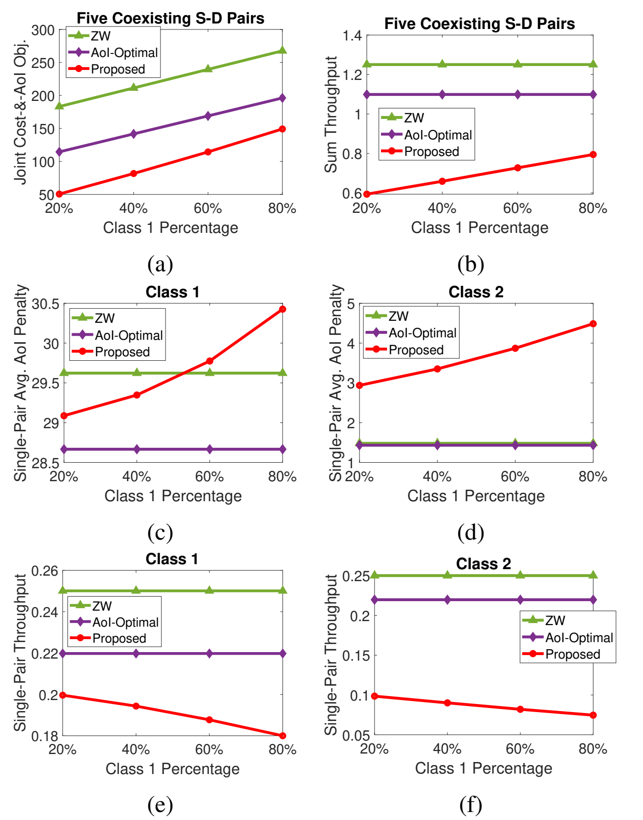

Fig. 4: Simulation results for five coexisting S-D pairs.

V. Conclusion

We have derived the optimal policy that jointly minimizes the sum of AoI penalties and the shared network cost across multiple coexisting traffics, while optimally balancing the heterogeneously timeliness requirements, heterogeneous throughput-to-cost relationships, and heterogeneous underlying delay distributions.

[4]Each S-D pair is allocated with less throughput when the Class-1 percentage increases, see Figs. 4e and 4f. But because each Class-1 pair has a higher throughput than that of a Class-2 pair and because we gradually replace the Class-2 pair(s) by Class-1 pair(s), the sum throughput still increases, see Fig. 4b.

References

[1] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Network*, vol. 32, no. 2, pp. 78–84, 2018.

[2] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 119–125, 2018.

[3] D. A. Chekired, M. A. Togou, L. Khoukhi, and A. Ksentini, "5G-slicing-enabled scalable SDN core network: Toward an ultra-low latency of autonomous driving service," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1769–1782, 2019.

[4] 3rd Generation Partnership Project (3GPP), "Release 16." [Online]. Available: www.3gpp.org/release-16

[5] X. Song and J. W.-S. Liu, "Performance of multiversion concurrency control algorithms in maintaining temporal consistency," in *Proceedings., Fourteenth Annual International Computer Software and Applications Conference*. IEEE, 1990, pp. 132–139.

[6] C. Kam, S. Kompella, and A. Ephremides, "Age of information under random updates," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 66–70.

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

[8] L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1681–1685.

[9] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of information in random access channels," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1770–1775.

[10] B. Buyukates, A. Soysal, and S. Ulukus, "Age of information in multihop multicast networks," *Journal of Communications and Networks*, vol. 21, no. 3, pp. 256–267, 2019.

[11] C.-H. Tsai and C.-C. Wang, "Unifying AoI minimization and remote estimation — optimal sensor/controller coordination with random two-way delay," in *2020 IEEE International Conference on Computer Communications (INFOCOM)*. IEEE, 2020, pp. 466–475.

[12] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 3008–3012.

[13] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.

[14] Y. Sun and B. Cyr, "Sampling for data freshness optimization: Non-linear age functions," *Journal of Communications and Networks*, vol. 21, no. 3, pp. 204–219, 2019.

[15] C.-H. Tsai and C.-C. Wang, "Age-of-information revisited: Two-way delay and distribution-oblivious online algorithm," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1782–1787.

[16] A.-H. El Fawal, A. Mansour, F. Le Roy, D. Le Jeune, and A. Hamié, "RACH overload congestion mechanism for M2M communication in LTE-A: Issues and approaches," in *2017 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2017, pp. 1–6.

[17] 3rd Generation Partnership Project (3GPP), "Release 15." [Online]. Available: www.3gpp.org/release-15

[18] H. Huang, D. Qiao, and M. C. Gursoy, "Age-energy tradeoff in fading channels with packet-based transmissions," *arXiv preprint arXiv:2005.05610*, 2020.

[19] E. Fountoulakis, N. Pappas, M. Codreanu, and A. Ephremides, "Optimal sampling cost in wireless networks with age of information constraints," *arXiv preprint arXiv:2003.02512*, 2020.

[20] J. Lou, X. Yuan, S. Kompella, and N.-F. Tzeng, "AoI and throughput tradeoffs in routing-aware multi-hop wireless networks," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 476–485.

[21] E. T. Ceran, D. Gündüz, and A. György, "Average age of information with hybrid ARQ under a resource constraint," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1900–1913, 2019.

[22] G. Yao, A. M. Bedewy, and N. B. Shroff, "Age minimization transmission scheduling over time-correlated fading channel under an average energy constraint," *arXiv preprint arXiv:2012.02958*, 2020.

[23] I. Kadota, A. Sinha, and E. Modiano, "Scheduling algorithms for optimizing age of information in wireless networks with throughput constraints," *IEEE/ACM Transactions on Networking*, vol. 27, no. 4, pp. 1359–1372, 2019.

[24] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 193–204, 2017.

[25] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor, "Timely estimation using coded quantized samples," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1812–1817.

[26] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 534–556, 2019.

[27] S. Farazi, A. G. Klein, and D. R. Brown, "Age of information in energy harvesting status update systems: When to preempt in service?" in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2436–2440.

[28] A. S. Tanenbaum, D. Wetherall *et al.*, "Computer networks," pp. I–XVII, 1996.

[29] R. Talak, S. Karaman, and E. Modiano, "Minimizing age-of-information in multi-hop wireless networks," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 486–493.

[30] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[31] L. Kleinrock, *Queueing systems, volume 2: Computer applications*. Wiley-Interscience, 1976.

[32] A. J. Coulson, A. G. Williamson, and R. G. Vaughan, "A statistical basis for lognormal shadowing effects in multipath fading channels," *IEEE Transactions on Communications*, vol. 46, no. 4, pp. 494–502, 1998.