2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, September 24-27, 2019

Timely Cloud Computing: Preemption and Waiting

Ahmed Arafa[1], Roy D. Yates[2], and H. Vincent Poor[1]

[1]Electrical Engineering Department, Princeton University
[2]Department of Electrical and Computer Engineering, Rutgers University

***Abstract*— The notion of timely status updating is investigated in the context of cloud computing. Measurements of a time-varying process of interest are acquired by a sensor node, and uploaded to a cloud server to undergo some required computations. These computations have random service times that are independent and identically distributed across different uploads. After the computations are done, the results are delivered to a monitor, constituting an *update*. The goal is to keep the monitor continuously fed with fresh updates over time, which is assessed by an *age-of-information* (AoI) metric. A scheduler is employed to optimize the measurement acquisition times. Following an update, an idle waiting period may be imposed by the scheduler before acquiring a new measurement. The scheduler also has the capability to *preempt* a measurement in progress if its service time grows above a certain *cutoff* time, and upload a fresher measurement in its place. Focusing on stationary deterministic policies, in which waiting times are deterministic functions of the instantaneous AoI and the cutoff time is fixed for all uploads, it is shown that the optimal waiting policy that minimizes the long term average AoI has a threshold structure, in which a new measurement is uploaded following an update only if the AoI grows above a certain threshold that is a function of the service time distribution and the cutoff time. The optimal cutoff is then found for standard and shifted exponential service times. While it has been previously reported that waiting before updating can be beneficial for AoI, it is shown in this work that preemption of *late* updates can be even more beneficial.**

I. Introduction

We consider the problem of timely computing. The setting is motivated by some applications in which monitoring a time-varying process of interest can be computationally demanding. Hence, instead of extracting useful information from local data measurements acquired by sensor nodes, measurements are uploaded to a cloud server that can handle heavy-duty computation tasks, and send them back in the form of updates. Computation times, however, are random, and the process may have already changed by the time they are done. We therefore investigate the benefits of preempting an upload in progress and replacing it by a new, *fresher*, one. Such freshness/timeliness is assessed by the *age-of-information* (AoI), defined as the time elapsed since the latest received update.

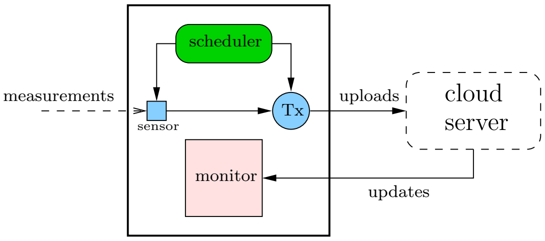

Fig. 1. A scheduler decides on when to acquire new measurements by the sensor and send them to the cloud server by the transmitter. The server updates the system's monitor after it completes the required computations.

Lots of work pertaining to AoI minimization have appeared in the recent literature, with frameworks that include queuing, optimization and scheduling, energy harvesting, remote estimation, and coding, see, e.g., [1]–[14]. Of particular relevance to our work are those in [15]–[23], which show that the notion of preemption of updates in service and replacing them by new ones is viable for AoI minimization in various settings. This is mainly owing to the nature of AoI that promotes sending fresh updates. This is discussed in a queuing framework in [15]–[17], and more recently in [18] that also extends to the case of multiple sources. Different from [15]–[18] that focus on exponential service times, the work in [19] considers general service time distributions for multiple Poisson sources with preemption. Preemption for general arrival and service time distributions, for a single source, has been recently studied in [20]. Reference [21] characterizes settings for which preemption is age-minimal, subject to energy harvesting constraints with Poisson arrivals (of both energy and updates) and exponential service times. The studies in [22], [23] investigate a similar tradeoff, under different system models, namely, that while preemption lets the system work with the freshest information, it leads to restarting service from the beginning. Thus, a decision has to be made on whether to drop the newly arriving updates or switch to them via preemption. Recently, in the context of computing, AoI analysis has been carried out through various tandem queuing models in [24]–[27], and through a task-specific age metric in [28]. The notion of sending timely measurements to the cloud has been discussed in the context of gaming in [29].

In this paper we investigate the tradeoff in [22], [23] in a cloud computing setting. Different from [22], [23], however, we consider a *generate-at-will* model, in which measurement times are controlled by a scheduler. Each measurement is uploaded to a cloud server to undergo some computations before being sent back as an update. *The scheduler has the ability to preempt a measurement in service if its computation time is larger than a certain cutoff time and upload a fresher one instead.* After an update is eventually received, the acquisition of a new measurement may be scheduled after an idle waiting period. We note that due to preemption, the optimal waiting policy derived in [3] does not apply in our setting.

This research was supported in part by the National Science Foundation under Grants CCF-0939370, CCF-1513915 and CCF-1717041.

978-1-7281-3151-1/19/$31.00 ©2019 IEEE

Focusing on stationary deterministic policies, in which cutoff times are constant and waiting times are function of the instantaneous AoI, we show that optimal waiting has a *threshold* structure. Specifically, a new measurement is uploaded, following an update, only if the AoI grows above a certain threshold that is a function of the cutoff time and the service time distribution. Such function is given in *closed-form*. We also provide a necessary and sufficient condition on the optimality of zero-wait policies, in which a new measurement is uploaded just-in-time as an update is received. When zero-wait is not optimal, we provide a a relatively simple method of evaluating the long term average AoI through a bisection search. We then discuss the evaluation of the optimal cutoff time explicitly under exponential service times. Finally, we compare the proposed preemption and waiting scheme to three baselines: no preemption and zero-waiting; no preemption and optimal waiting, the scheme proposed in [3]; and optimal preemption and zero-waiting. While it is demonstrated that our proposed scheme perfoms best, our results also show that, depending on the system parameters, the optimal preemption and zero-waiting policy can actually beat the no preemption and optimal waiting policy. This sheds light on the fact that, in some situations, working with fresh measurements provides the highest gains in terms of AoI.

II. System Model and Problem Formulation

We consider a system comprised of a sensor, a scheduler, a transmitter, a cloud computing server and a monitor. The overall goal is to keep the system's monitor continuously fed with *fresh* status updates pertaining to a physical phenomenon of interest. Such updates, however, require some heavy-duty computations on the raw data measurements acquired by the sensor that need to be carried out by the cloud computing server. Therefore, in order for a status update to reach the monitor, the following series of events need to occur. First, the scheduler decides on when to acquire a new data measurement by the sensor, and send (upload) it to the server by the transmitter. The server then undertakes the computations and feeds back the end result to the monitor in the form of an *update*. Hence, the goal is to design a scheduling policy such that these updates reach the monitor in a timely manner. A depiction of the system model considered is shown in Fig. 1. Hereafter, we will refer to data sent to the server by uploads, and data received from the server by updates.

Uploads are time-stamped so that when updates eventually reach the monitor, the system knows when their corresponding measurements were acquired. We use an AoI metric to assess the timeliness of updates at the monitor. This is defined as

$$a(t) = t - u(t), \qquad (1)$$

where $u(t)$ is the time stamp of the latest update that has reached the monitor.

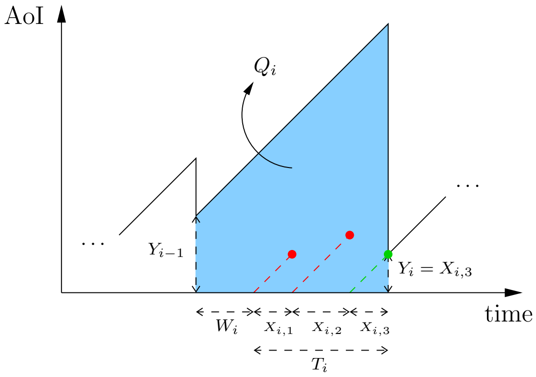

Fig. 2. AoI evolution example in the ith epoch. Red lines denote preemptions and the green line denotes completed service. In this example $N_i = 3$.

To minimize the AoI, measurements are uploaded to the cloud server right away after being acquired by the sensor. We assume that upload transmission times are negligible. However, each measurement consumes a computational time at the cloud server denoted as the service time. Service times of different measurements are independent and identically distributed (i.i.d.) according to the distribution of a random variable X. Depending on the application or the task being considered, the server may incur a constant delay before the actual computation starts. Let us denote such time by $c \in \mathbb{R}_+$, which is the largest constant such that

$$X \geq c \quad \text{a.s.} \qquad (2)$$

This is without loss of generality since $X \geq 0$ a.s.[1] Motivated by freshness, the scheduler is capable of *preempting* the current upload in service if its service time surpasses a certain *cutoff* time. Thus, an update will reach the monitor only if its service ends within the cutoff time. Following a preemption, a new measurement is taken and uploaded immediately. Following an update delivery, however, the scheduler may *wait* for some idle time before uploading a new measurement.

We denote by the ith epoch the time in between the reception of the $(i-1)$th and the ith updates. The ith epoch starts with age Y_{i-1} and ends with age Y_i.[2] A waiting period of W_i time units occurs at the start of the epoch, through which the server is idle. After that, the first measurement in the epoch is acquired and uploaded to the server. Note that, depending on the preemption policy, there can be multiple uploads during a single epoch. We denote by N_i the number of uploads during the ith epoch, with $N_i \geq 1$, and $N_i - 1$ denoting the number of preemptions. Let $X_{i,j}$ denote the service time of the jth upload during the ith epoch, $1 \leq j \leq N_i$. Note that $X_{i,j}$'s are i.i.d $\sim X$. Now observe that only the X_{i,N_i} service time period concludes with an update sent back to the monitor, and all other service time periods end with a preemption. Therefore, the ith epoch ends with age

$$Y_i = X_{i,N_i}. \qquad (3)$$

We denote by T_i the server's busy period in the ith epoch, defined as

$$T_i \triangleq X_{i,1} + X_{i,2} + \cdots + X_{i,N_i}. \qquad (4)$$

[1] One might consider the constant c a necessary overhead to initiate service at the cloud server for each measurement.

[2] We assume that the first epoch starts with some given age Y_0 at time 0.

Lastly, let L_i denote the ith epoch length given by

$$L_i = W_i + T_i. \tag{5}$$

In Fig. 2, we show an example sample path of how the AoI may evolve during the ith epoch. From the figure, the area under the AoI curve during the ith epoch, Q_i, is given by

$$Q_i = Y_{i-1}L_i + \frac{1}{2}L_i^2. \tag{6}$$

We are interested in minimizing the long term average AoI. It is clear that such quantity depends on the choices of the cutoff and waiting times that the scheduler makes. Let $\gamma_{i,j}$ denote the cutoff time after which the jth upload in the ith epoch is preempted. In other words, given $\gamma_{i,j}$, the scheduler preempts the jth upload in the ith epoch if $X_{i,j}$ grows above $\gamma_{i,j}$ time units. Clearly,

$$\gamma_{i,j} \geq c \tag{7}$$

must hold $\forall i, j$ in view of (2). The set $\{\gamma_{i,j}\}$ now constitutes a *cutoff policy*, while the set $\{W_i\}$ denotes a *waiting policy*. Let π denote a scheduling policy $\{W_i, \gamma_{i,j}\}$. The goal is to solve the following problem:

$$\min_{\pi} \limsup_{n\to\infty} \frac{\sum_{i=1}^{n} \mathbb{E}\left[Q_i\right]}{\sum_{i=1}^{n} \mathbb{E}\left[L_i\right]}, \tag{8}$$

where $\mathbb{E}\left[\cdot\right]$ is the expectation operator.

III. Stationary Deterministic Policies

Observe that the optimal policy π^* that solves problem (8) may be such that the waiting and cutoff times of the ith epoch depend on the history of events, e.g., AoI evolution, number of preemptions, service time realizations, before, and during, the ith epoch. To alleviate the difficulty of tracking all such history, and motivated by the fact that service times are i.i.d., we focus on policies that are characterized by the following two main features: 1) the waiting time W_i in the ith epoch is given by a *deterministic* function of the starting AoI Y_{i-1},

$$W_i \triangleq w\left(Y_{i-1}\right), \tag{9}$$

for some function $w : \mathbb{R}_+ \to \mathbb{R}_+$; and 2) the cutoff times $\{\gamma_{i,j}\}$ in the ith epoch are given by *deterministic* functions of the instantaneous AoI,

$$\gamma_{i,j} \triangleq \gamma_j\left(a_{i,j}\right), \tag{10}$$

for some function $\gamma_j : \mathbb{R}_+ \to [c, \infty]$, $\forall j$, with $a_{i,j}$ denoting the AoI just before the jth upload occurs in the ith epoch.

Let Π_s denote the set of policies that abide by the above structure. Note that any $\pi \in \Pi_s$ induces *stationary* distributions $Q_i \sim Q$ and $L_i \sim L$ for all epochs. Therefore, under Π_s, problem (8) reduces to

$$\min_{\pi\in\Pi_s} \frac{\mathbb{E}\left[Q\right]}{\mathbb{E}\left[L\right]}. \tag{11}$$

Problem (11) is an optimization problem over a single epoch. In the sequel, we drop the index i for convenience. We now have the following lemma:

Lemma 1 *In the optimal solution of problem (11), all cutoff functions are equivalent. That is,*

$$\gamma_j\left(a_j\right) \equiv \gamma\left(a_j\right), \quad \forall j, \tag{12}$$

for some $\gamma : \mathbb{R}_+ \to [c, \infty]$.

Proof: Let the optimal cutoff function $\gamma_1(\cdot)$ be given. Note that the system is idle before the first upload. Thus, $\gamma_1(a_1)$ represents the optimal cutoff time for the AoI to evolve starting from an idle state at age a_1. Now assume that the first upload is preempted after $\gamma_1(a_1)$, whence the age becomes $a_2 = a_1 + \gamma_1(a_1)$. Observe that the system becomes *instantly* idle right before the second upload occurs. Since service times are i.i.d., therefore $\gamma_2(a_2)$ should also represent the optimal cutoff time for the AoI to evolve starting from an idle state at age a_2. This shows that $\gamma_2(a_2) = \gamma_1(a_2)$ must hold, otherwise $\gamma_1(\cdot)$ would not be optimal. Similar arguments follow for $\gamma_j(\cdot)$, $j \geq 3$. Therefore, all cutoff functions are equivalent. ■

In the sequel, we further focus on the case in which the cutoff function $\gamma(\cdot)$ is a constant. That is, with a slight abuse of notation,

$$\gamma\left(a_j\right) = \gamma, \quad \forall j, \tag{13}$$

for some $\gamma \geq c$. We call this the γ-*cutoff policy*. Considering such policy is motivated by the fact that service times are i.i.d.; it also sets a fixed maximum value of γ on the starting AoI of each epoch.

Now let the following quantities be (re)defined for the epoch in consideration: $\overline{Y}$ is the starting AoI; $W = w\left(\overline{Y}\right)$ is the waiting time after it starts; T is the server's busy period; X_j is the jth upload service time; N is the total number of uploads; and $\underline{Y}$ is the ending AoI. Observe that under a γ-cutoff policy, given $N = n$, $X_1 = \cdots = X_{n-1} = \gamma$ and $X_n = \underline{Y} \leq \gamma$ a.s. Also observe that the function $w(\cdot)$ is now restricted to the domain $[0, \gamma]$, and that $\overline{Y}$ and $\underline{Y}$ are i.i.d $\sim Y$. To evaluate the distribution of the age Y, let us define $p \triangleq \mathbb{P}\left(X \leq \gamma\right)$, where $\mathbb{P}\left(\cdot\right)$ is the probability measure. Therefore the probability distribution function (PDF) of Y is given by

$$f_Y(y) = \begin{cases} \frac{f_X(y)}{p}, & c \leq y \leq \gamma \\ 0, & \text{otherwise} \end{cases}, \tag{14}$$

where $f_X(\cdot)$ denotes the PDF of the service time X.[3]

We note that problem (11) is structurally different from the setting considered in [3]. There, an epoch could only consist of one packet in service until it finishes, and hence the AoI at the end of the epoch relates to that packet's acquisition time. In our setting, owing to the preemption capability, there can be multiple uploads in a single epoch, and hence the AoI at the end of the epoch does not necessarily relate to the first upload time. The optimal waiting policy derived in [3], therefore, does not apply in our setting.

[3] We focus on continuous random variables, and assume that γ and the distribution of X are such that $p > 0$.

Solving problem (11) is tantamount to characterizing the optimal waiting function $w^*(\cdot)$ and the optimal cutoff time γ^*. In the next sections, we do so sequentially as follows: we first characterize $w^*(\cdot)$ for a fixed value of γ, and then we determine γ^* for specific service time distributions.

IV. Threshold Waiting Policy

In this section, we evaluate the optimal waiting policy for fixed cutoff time γ. The main result is that the optimal waiting policy exhibits a threshold structure, in which a new upload occurs only if the AoI grows above a certain threshold that depends on the service time distribution and the fixed cutoff time. Toward showing that, we need to evaluate some expressions first. We start with

$$\mathbb{P}(N=n)=(1-p)^{n-1}p, \quad n\geq 1, \tag{15}$$

i.e., N is a geometric random variable with parameter p. It is useful to note that $\mathbb{E}[N]=\frac{1}{p}$ and $\mathbb{E}\left[N^2\right]=\frac{2-p}{p^2}$. Using iterated expectations, we now have

$$\begin{aligned}\mathbb{E}[T]=&\sum_{n=1}^{\infty}\mathbb{P}(N=n)\,\mathbb{E}[X_1+X_2+\cdots+X_n]\\=&\sum_{n=1}^{\infty}\mathbb{P}(N=n)\left((n-1)\gamma+\mathbb{E}[\underline{Y}]\right)\\=&\left(\frac{1}{p}-1\right)\gamma+\mathbb{E}[\underline{Y}].\end{aligned} \tag{16}$$

Thus, the expected epoch length is given by

$$\mathbb{E}[L]=\mathbb{E}\left[w\left(\overline{Y}\right)\right]+\mathbb{E}[T], \tag{17}$$

with $\mathbb{E}[T]$ given by (16). Proceeding similarly, we have

$$\begin{aligned}\mathbb{E}\left[T^2\right]=&\sum_{n=1}^{\infty}\mathbb{P}(N=n)\,\mathbb{E}\left[(X_1+X_2+\cdots+X_n)^2\right]\\=&\sum_{n=1}^{\infty}\mathbb{P}(N=n)\left((n-1)^2\gamma^2+2(n-1)\gamma\mathbb{E}[\underline{Y}]+\mathbb{E}\left[\underline{Y}^2\right]\right)\\=&\left(\frac{2-p}{p^2}-\frac{2}{p}+1\right)\gamma^2+2\left(\frac{1}{p}-1\right)\gamma\mathbb{E}[\underline{Y}]+\mathbb{E}\left[\underline{Y}^2\right].\end{aligned} \tag{18}$$

We now have

$$\begin{aligned}\mathbb{E}[Q]=&\mathbb{E}\left[\overline{Y}\left(w\left(\overline{Y}\right)+T\right)\right]+\frac{1}{2}\mathbb{E}\left[\left(w\left(\overline{Y}\right)+T\right)^2\right]\\=&\mathbb{E}\left[\overline{Y}w\left(\overline{Y}\right)\right]+\mathbb{E}\left[\overline{Y}\right]\mathbb{E}[T]+\frac{1}{2}\mathbb{E}\left[w^2\left(\overline{Y}\right)\right]\\&+\mathbb{E}\left[w\left(\overline{Y}\right)\right]\mathbb{E}[T]+\frac{1}{2}\mathbb{E}\left[T^2\right],\end{aligned} \tag{19}$$

with $\mathbb{E}[T]$ and $\mathbb{E}\left[T^2\right]$ given by (16) and (18), respectively, and the second equality follows by independence of $\overline{Y}$ and T.

To find the optimal $w^*(\cdot)$, we need to solve the following functional optimization problem:

$$\begin{aligned}\min_{w(\cdot)}\quad&\frac{\mathbb{E}[Q]}{\mathbb{E}[L]}\\\text{s.t.}\quad&w(t)\geq 0,\quad c\leq t\leq\gamma.\end{aligned} \tag{20}$$

To solve the above problem, we follow Dinkelbach's approach [30] and introduce the following auxiliary problem for some fixed parameter $\lambda\geq 0$:

$$\begin{aligned}g(\lambda)\triangleq\min_{w(\cdot)}\quad&\mathbb{E}[Q]-\lambda\mathbb{E}[L]\\\text{s.t.}\quad&w(t)\geq 0,\quad c\leq t\leq\gamma.\end{aligned} \tag{21}$$

One can show that $g(\lambda)$ is decreasing in λ, and that the optimal solution of problem (20) is given by λ^* that solves $g(\lambda^*)=0$ [30]. By monotonicity of $g(\cdot)$, λ^* can be found by, e.g., a bisection search. Focusing on problem (21), we introduce the following Lagrangian [31]:

$$\mathcal{L}=\mathbb{E}[Q]-\lambda\mathbb{E}[L]-\int_c^{\gamma}w(\tau)\eta(\tau)d\tau, \tag{22}$$

where $\eta(\cdot)$ is a Lagrange multiplier. Substituting (17) and (19) above, and after some rearrangements we get

$$\begin{aligned}\mathcal{L}=&\int_c^{\gamma}\left(\left(\tau+\mathbb{E}[T]-\lambda\right)w(\tau)+\frac{1}{2}w^2(\tau)\right)f_Y(\tau)d\tau\\&+\mathbb{E}\left[\overline{Y}\right]\mathbb{E}[T]+\frac{1}{2}\mathbb{E}\left[T^2\right]-\lambda\mathbb{E}[T]-\int_c^{\gamma}w(\tau)\eta(\tau)d\tau.\end{aligned} \tag{23}$$

Now taking the (functional) derivative of $\mathcal{L}$ with respect to $w(t)$, $c\leq t\leq\gamma$, and equating to 0 we have

$$\left(t+\mathbb{E}[T]-\lambda+w^*(t)\right)f_Y(t)-\eta(t)=0. \tag{24}$$

Rearranging the above, we get that

$$w^*(t)=\lambda-\mathbb{E}[T]-t+\frac{\eta(t)}{f_Y(t)}. \tag{25}$$

We note that there are different methods through which one can conclude that the optimal waiting policy satisfies (25). These are discussed in Appendix D for completeness. Now using complementary slackness [31], (25) further gives

$$w^*(t)=\left[\lambda-\mathbb{E}[T]-t\right]^+,\quad c\leq t\leq\gamma, \tag{26}$$

where $[\cdot]^+\triangleq\max(\cdot,0)$. This makes the AoI right after the waiting period, when the first measurement in the epoch gets uploaded, equal to

$$t+w^*(t)=\max\{t,\lambda-\mathbb{E}[T]\}, \tag{27}$$

which comes directly from the fact that $w^*(t)>0$ if and only if (iff) $\lambda-\mathbb{E}[T]>t$. Observe that the value of t, the realization of $\overline{Y}$, represents the AoI at the beginning of the epoch. Hence, one could interpret the optimal waiting policy as a *threshold* policy, in which the first measurement in the epoch gets uploaded only if the AoI grows above $\lambda-\mathbb{E}[T]$.

To have an operational significance, however, the threshold $\lambda-\mathbb{E}[T]$ must be positive. The next lemma verifies that this is indeed the case. The proof is in Appendix A.

Lemma 2 *The optimal solution of problem (20), λ^*, satisfies $\lambda^*>\mathbb{E}[T]$.*

Observe that while Lemma 2 shows that the threshold

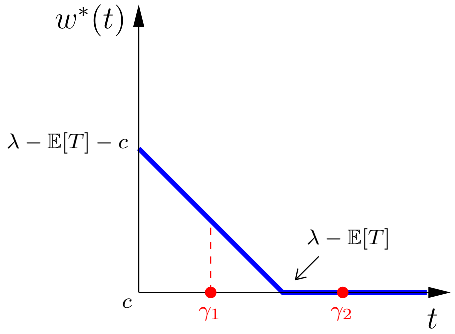

Fig. 3. The optimal waiting policy versus time. We show two example choices of γ in red. For γ_1, we always wait before uploading a new measurement following an update, while for γ_2 it depends on the value of t. Lemma 4 shows that the situation of γ_1 cannot be optimal.

is positive, a zero-wait policy can still be optimal if the threshold's value is no larger than c. The next lemma quantifies this relationship. The proof is in Appendix B.

Lemma 3 *A zero-wait policy, in which $w^*(t) = 0$, $\forall t \in [c, \gamma]$, is optimal for problem (20) iff*

$$\frac{\frac{1}{2}\left(\frac{1}{p} - 1\right)\gamma^2 + \frac{1}{2}\mathbb{E}\left[Y^2\right]}{\left(\frac{1}{p} - 1\right)\gamma + \mathbb{E}\left[Y\right]} \leq c. \tag{28}$$

The optimal AoI under a zero-wait policy is directly given by substituting $w^*(t) = 0$, $\forall t$ in (17) and (19) to get

$$\begin{aligned} \lambda^*_{zw} &= \frac{\mathbb{E}\left[Q\right]}{\mathbb{E}\left[L\right]} \\ &= \frac{\mathbb{E}\left[\overline{Y}\right]\mathbb{E}\left[T\right] + \frac{1}{2}\mathbb{E}\left[T^2\right]}{\mathbb{E}\left[T\right]} \\ &= \mathbb{E}\left[Y\right] + \frac{\frac{1}{2}\mathbb{E}\left[T^2\right]}{\mathbb{E}\left[T\right]}, \end{aligned} \tag{29}$$

where the subscript zw stands for zero-wait.

Now that we established a necessary and sufficient condition for the optimality of a zero-wait policy in Lemma 3, we proceed by investigating the case in which the inequality condition in (28) does *not* hold. First, an immediate corollary follows in this case.

Corollary 1 *The optimal solution of problem (20), λ^*, satisfies $\lambda^* > \mathbb{E}\left[T\right] + c$ iff (28) does not hold.*

Now observe that for $\gamma < \lambda^* - \mathbb{E}\left[T\right]$, one would *always* wait before uploading a new measurement following an update, and that for $\gamma \geq \lambda^* - \mathbb{E}\left[T\right]$, it depends on the realization of $\overline{Y}$ (the value of t) as indicated in (26). We illustrate this behavior in Fig. 3, and settle this issue in the next lemma by showing that the situation of γ_1 in Fig. 3 *cannot* be optimal.[4] We note that the result of the lemma holds regardless of whether (28) holds or not. The proof is in Appendix C.

Lemma 4 *The optimal solution of problem (20), λ^*, satisfies $\gamma \geq \lambda^* - \mathbb{E}\left[T\right]$.*

In summary, to find the optimal AoI for fixed γ one should start by examining (28). If it holds, then $\lambda^* = \lambda^*_{zw}$ in (29). Else, using Corollary 1 and Lemma 4, one has the following bounds on the optimal AoI:

$$\mathbb{E}\left[T\right] + c < \lambda^* \leq \mathbb{E}\left[T\right] + \gamma, \tag{30}$$

which facilitates evaluating λ^* that solves $g(\lambda^*) = 0$ using a bisection search in the interval $(\mathbb{E}\left[T\right] + c, \mathbb{E}\left[T\right] + \gamma]$.

Now it remains to choose the best γ that minimizes λ^*. We discuss this in the next section.

V. Optimal γ-Cutoff Policy

It is not direct to get a closed-form expression of the optimal λ^* in terms of γ for general service time distributions. In fact, even for specific distributions this can also be a difficult task. In this section, our goal is to provide some insight on how the service time distribution can affect the choice of the optimal cutoff γ^*. To avoid confusion, let the optimal AoI as a function of the cutoff value, derived in Section IV, be denoted by $\lambda^*\left(\gamma\right)$, and define $\lambda^{**} \triangleq \lambda^*\left(\gamma^*\right)$. Our approach will be as follows: we will first fix $\gamma \geq c$ and evaluate $\lambda^*\left(\gamma\right)$ as discussed toward the end of Section IV; and then we will evaluate γ^* that minimizes $\lambda^*\left(\gamma\right)$, i.e., achieves λ^{**}, numerically.

We will consider an exponential service time distribution with $c = 0$ along with its shifted version with $c > 0$. Clearly, the zero-wait policy is not optimal for distributions with $c = 0$, as inferred from the inequality (28). In this case, $\lambda^*\left(\gamma\right)$ can be evaluated by a bisection search using the bounds in (30). On the other hand, for $c > 0$, $\lambda^*\left(\gamma\right)$ is given in closed-form by λ^*_{zw} in (29) for values of γ that satisfy (28), and is evaluated by a bisection search using the bounds in (30) otherwise. As we will see, in some situations evaluating γ^* will be a direct consequence of evaluating the bounds in (30).

A. Standard Exponential

Let $X \sim \exp(1)$.[5] Since $c = 0$, we aim at evaluating the bounds in (30). Toward that, one can directly compute the following quantities:

$$p = 1 - e^{-\gamma}, \tag{31}$$

$$\mathbb{E}\left[Y\right] = \frac{1 - \left(1 + \gamma\right)e^{-\gamma}}{1 - e^{-\gamma}}. \tag{32}$$

This directly gives $\mathbb{E}\left[T\right] = 1$, and hence

$$1 < \lambda^*\left(\gamma\right) \leq 1 + \gamma, \tag{33}$$

upon which one can see that γ^* is infinitesimal. As mentioned before, this is one instance where evaluating the bounds in (30) directly gives γ^*. Therefore, in this case, λ^{**} can be made arbitrarily close to 1 by choosing γ^* arbitrarily close to 0.

[4] We note that Fig. 3 is only explanatory and that in reality the choice of γ also affects the values of λ^* and $\mathbb{E}\left[T\right]$.

[5] One can always choose a time unit such that the service rate is unity.

B. Shifted Exponential

We now focus on the shifted version of the above, in which

$$f_X(x) = e^{-(x-c)}, \quad x \geq c > 0. \tag{34}$$

Based on this, for $\gamma \geq c$, one can directly compute

$$p = 1 - e^{-(\gamma-c)}, \tag{35}$$

$$\mathbb{E}[Y] = \frac{1 + c - (1+\gamma)\, e^{-(\gamma-c)}}{1 - e^{-(\gamma-c)}}, \tag{36}$$

$$\mathbb{E}\left[Y^2\right] = \frac{2 + 2c + c^2 - \left(2 + 2\gamma + \gamma^2\right) e^{-(\gamma-c)}}{1 - e^{-(\gamma-c)}}. \tag{37}$$

Upon substituting all the above in (28) and simplifying, we get that the zero-wait policy is optimal iff

$$1 - \frac{1}{2}c^2 \leq (1 + \gamma - c)\, e^{-(\gamma-c)}. \tag{38}$$

Observe that the above is satisfied for all values of $\gamma \geq c$ if $c \geq \sqrt{2}$. Next, note that $(1+\gamma-c)\, e^{-(\gamma-c)}$ is decreasing in γ, and has a maximum value of 1 when $\gamma = c$. This shows that there exists a unique $\bar{\gamma}(c) > c$ that satisfies the above inequality with equality if $c < \sqrt{2}$. Thus, the inequality is satisfied for $c < \sqrt{2}$ iff $\gamma \leq \bar{\gamma}(c)$. Based on the above, the zero-wait policy is optimal in the following cases: 1) $c \geq \sqrt{2}$, or 2) $c < \sqrt{2}$ and $\gamma \leq \bar{\gamma}(c)$. On the other hand the zero-wait policy is not optimal if $c < \sqrt{2}$ and $\gamma > \bar{\gamma}(c)$.

In Fig. 4, we plot the optimal cutoff γ^* and the corresponding AoI λ^* versus c. We also show $\bar{\gamma}(c)$ on the figure to indicate whether zero-wait is optimal for $c < \sqrt{2}$. We see from the figure that the zero-wait policy is *not* optimal for all values of $c < \sqrt{2}$ since $\gamma^* > \bar{\gamma}(c)$; it is only optimal for $c \geq \sqrt{2}$. Note that $\bar{\gamma}(c)$ is not defined (and not needed) for $c \geq \sqrt{2}$, and is therefore not shown on the figure.

In Fig. 5, we compare the optimal policy derived in this paper to other benchmarks. The first is the vanilla version of status updating, denoted *no cutoff & zero-wait*, in which an upload is never preempted, and a new upload takes place once an update is received. The second is also a zero-wait policy yet with optimizing the cutoff value, denoted *optimal cutoff & zero-wait*. The third is that of [3], denoted *no cutoff & optimal wait*, in which the waiting time is optimized and uploads are never preempted. We see that our policy beats all benchmarks, especially for small values of c. Another interesting note is that for for $c \lessapprox 0.25$, optimizing the cutoff turns out to be better, age-wise, than optimizing the waiting time. Indeed, the *optimal cutoff & zero-wait* policy beats the *no cutoff & optimal wait* policy of [3] for $c \lessapprox 0.25$.

VI. Conclusion

A cloud computing status updating system has been considered, in which computations are carried out on raw data measurements uploaded to a cloud server, and then returned in the form of updates to a monitor. Using an AoI metric, it has been shown that preemption of late updates, whose service times exceed a certain cutoff time, and replacing them by fresher measurements can enhance the overall AoI. Further, it has been shown that it is optimal to upload a new measurement to the server following an update only if the AoI grows above a certain threshold. Implications of such preemption and waiting policies have been discussed for exponential service time distributions, along with comparison with other benchmarks.

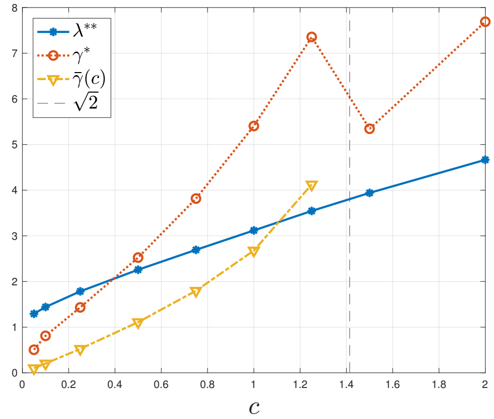

Fig. 4. Optimal AoI and cutoff values versus c for exponential service times. The vertical line denotes the critical value of $c = \sqrt{2}$, after which the zero-wait policy is optimal.

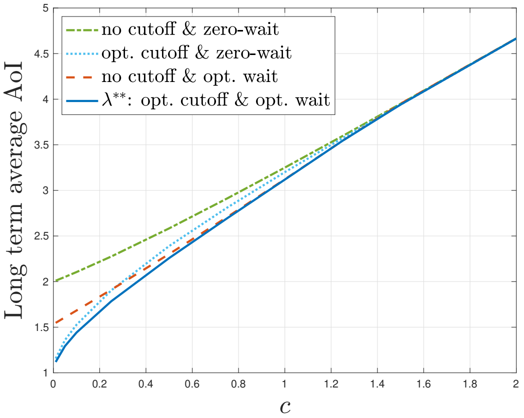

Fig. 5. Comparing the optimal policy to other bench marks versus c for exponential service times.

Appendix

A. Proof of Lemma 2

We show this by contradiction. Assume that $\lambda^* \leq \mathbb{E}[T]$. Then this would necessarily mean that $w^*(t) = 0$, $\forall t$, and hence (cf. (29))

$$\begin{aligned}\lambda^* &= \frac{\mathbb{E}\left[Y\right]\mathbb{E}[T] + \frac{1}{2}\mathbb{E}\left[T^2\right]}{\mathbb{E}[T]} \\ &= \mathbb{E}[T] - \left(\frac{1}{p} - 1\right)\gamma + \frac{\frac{1}{2}\mathbb{E}\left[T^2\right]}{\mathbb{E}[T]},\end{aligned} \tag{39}$$

where (39) follows by (16). Now for λ^* to be no larger than $\mathbb{E}[T]$, it must hold that

$$\frac{\frac{1}{2}\mathbb{E}\left[T^2\right]}{\mathbb{E}[T]} \leq \left(\frac{1}{p} - 1\right)\gamma. \tag{40}$$

Using (16) and (18), the above is tantamount to having

$$\frac{1}{2}\left(\frac{2-p}{p^2}-\frac{2}{p}+1\right)\gamma^2+\left(\frac{1}{p}-1\right)\gamma\mathbb{E}\left[\underline{Y}\right]+\frac{1}{2}\mathbb{E}\left[\underline{Y}^2\right]$$
$$\leq\left(\frac{1}{p}-1\right)^2\gamma^2+\left(\frac{1}{p}-1\right)\gamma\mathbb{E}\left[\underline{Y}\right], \quad (41)$$

which, upon some direct algebraic rearrangements, is equivalent to having

$$\frac{1}{2}\left(\frac{1}{p}-1\right)\gamma^2+\frac{1}{2}\mathbb{E}\left[\underline{Y}^2\right]\leq 0, \quad (42)$$

which is a clear contradiction.

B. Proof of Lemma 3

In view of (26), a zero-wait policy is optimal iff $\lambda^* \leq \mathbb{E}\left[T\right]+c$. Proceeding as in Appendix A, this is tantamount to adding c to the right hand side (RHS) of (40), or equivalently adding $c\mathbb{E}\left[T\right]$ to the RHSs of (41) and (42). Thus, a zero-wait policy is optimal iff

$$\frac{\frac{1}{2}\left(\frac{1}{p}-1\right)\gamma^2+\frac{1}{2}\mathbb{E}\left[\underline{Y}^2\right]}{\mathbb{E}\left[T\right]}\leq c. \quad (43)$$

Substituting (16) above directly gives (28).

C. Proof of Lemma 4

First, if (28) holds, then by Corollary 1 $\lambda^* \leq \mathbb{E}\left[T\right]+c \leq \mathbb{E}\left[T\right]+\gamma$.

We now show the result of the lemma when (28) does not hold. We show this by contradiction. Assume that $\gamma < \lambda^* - \mathbb{E}\left[T\right]$. Under that assumption, it holds by (26) that

$$w^*(t)=\lambda-\mathbb{E}\left[T\right]-t, \quad c\leq t\leq\gamma, \quad (44)$$

i.e., $w^*(t)>0$, $\forall t\in[c,\gamma]$. Therefore,

$$\mathbb{E}\left[w\left(\overline{Y}\right)\right]=\int_c^\gamma\left(\lambda-\mathbb{E}\left[T\right]-\tau\right)f_Y(\tau)d\tau$$
$$=\lambda-\mathbb{E}\left[T\right]-\mathbb{E}\left[\overline{Y}\right]. \quad (45)$$

Our goal now is to evaluate the value of λ^* by solving $g(\lambda^*)=0$, and show that it cannot be larger than $\mathbb{E}\left[T\right]+\gamma$, thereby reaching a contradiction. Toward that, we start by using the above to evaluate

$$\mathbb{E}\left[L\right]=\mathbb{E}\left[w\left(\overline{Y}\right)\right]+\mathbb{E}\left[T\right]=\lambda-\mathbb{E}\left[\overline{Y}\right]. \quad (46)$$

Since $\mathbb{E}\left[L\right]\geq 0$, it must hold that the optimal λ^* satisfies

$$\lambda^*\geq\mathbb{E}\left[\overline{Y}\right]. \quad (47)$$

This simple observation will prove to be useful later on.

Next, we have

$$\mathbb{E}\left[\overline{Y}w\left(\overline{Y}\right)\right]=\int_c^\gamma\tau\left(\lambda-\mathbb{E}\left[T\right]-\tau\right)\frac{f_Y(\tau)}{p}d\tau$$
$$=\left(\lambda-\mathbb{E}\left[T\right]\right)\mathbb{E}\left[\overline{Y}\right]-\mathbb{E}\left[\overline{Y}^2\right], \quad (48)$$

and

$$\mathbb{E}\left[w^2\left(\overline{Y}\right)\right]=\int_c^\gamma\left(\lambda-\mathbb{E}\left[T\right]-\tau\right)^2\frac{f_Y(\tau)}{p}d\tau$$
$$=\left(\lambda-\mathbb{E}\left[T\right]\right)^2-2\left(\lambda-\mathbb{E}\left[T\right]\right)\mathbb{E}\left[\overline{Y}\right]+\mathbb{E}\left[\overline{Y}^2\right]. \quad (49)$$

Substituting (45), (48) and (49) in (19) we get

$$\mathbb{E}\left[Q\right]=\left(\lambda-\mathbb{E}\left[T\right]\right)\mathbb{E}\left[\overline{Y}\right]-\mathbb{E}\left[\overline{Y}^2\right]+\mathbb{E}\left[\overline{Y}\right]\mathbb{E}\left[T\right]$$
$$+\frac{1}{2}\left(\lambda-\mathbb{E}\left[T\right]\right)^2-\left(\lambda-\mathbb{E}\left[T\right]\right)\mathbb{E}\left[\overline{Y}\right]+\frac{1}{2}\mathbb{E}\left[\overline{Y}^2\right]$$
$$+\left(\lambda-\mathbb{E}\left[T\right]-\mathbb{E}\left[\overline{Y}\right]\right)\mathbb{E}\left[T\right]+\frac{1}{2}\mathbb{E}\left[T^2\right]$$
$$=-\frac{1}{2}\mathbb{E}\left[\overline{Y}^2\right]+\frac{1}{2}\left(\lambda-\mathbb{E}\left[T\right]\right)^2$$
$$+\left(\lambda-\mathbb{E}\left[T\right]\right)\mathbb{E}\left[T\right]+\frac{1}{2}\mathbb{E}\left[T^2\right]$$
$$=\frac{1}{2}\lambda^2+\frac{1}{2}\mathbb{E}\left[T^2\right]-\frac{1}{2}\left(\mathbb{E}\left[T\right]\right)^2-\frac{1}{2}\mathbb{E}\left[\overline{Y}^2\right]. \quad (50)$$

The above can be further simplified by noting that using (16) and (18) we have

$$\mathbb{E}\left[T^2\right]-\left(\mathbb{E}\left[T\right]\right)^2$$
$$=\left(\frac{2}{p}-1\right)\left(\frac{1}{p}-1\right)\gamma^2+2\left(\frac{1}{p}-1\right)\gamma\mathbb{E}\left[\overline{Y}\right]+\mathbb{E}\left[\overline{Y}^2\right]$$
$$-\left(\frac{1}{p}-1\right)^2\gamma^2-2\left(\frac{1}{p}-1\right)\gamma\mathbb{E}\left[\overline{Y}\right]-\left(\mathbb{E}\left[\overline{Y}\right]\right)^2$$
$$=\frac{1}{p}\left(\frac{1}{p}-1\right)\gamma^2+\mathbb{E}\left[\overline{Y}^2\right]-\left(\mathbb{E}\left[\overline{Y}\right]\right)^2$$
$$=\frac{1-p}{p^2}\gamma^2+\mathbb{E}\left[\overline{Y}^2\right]-\left(\mathbb{E}\left[\overline{Y}\right]\right)^2, \quad (51)$$

which, upon substituting in (50) finally gives

$$\mathbb{E}\left[Q\right]=\frac{1}{2}\lambda^2+\frac{1-p}{2p^2}\gamma^2-\frac{1}{2}\left(\mathbb{E}\left[\overline{Y}\right]\right)^2. \quad (52)$$

Now using (46) and (52) we have

$$g(\lambda)=\mathbb{E}\left[Q\right]-\lambda\mathbb{E}\left[L\right]$$
$$=-\frac{1}{2}\lambda^2+\frac{1-p}{2p^2}\gamma^2-\frac{1}{2}\left(\mathbb{E}\left[\overline{Y}\right]\right)^2+\lambda\mathbb{E}\left[\overline{Y}\right]. \quad (53)$$

Thus, solving $g(\lambda^*)=0$ is equivalent to solving

$$\left(\lambda^*\right)^2-2\mathbb{E}\left[\overline{Y}\right]\lambda^*+\left(\mathbb{E}\left[\overline{Y}\right]\right)^2-\frac{1-p}{p^2}\gamma^2=0. \quad (54)$$

The above equation has two solutions, but only one of them is valid due to the inequality in (47). This is given by

$$\lambda^*=\mathbb{E}\left[\overline{Y}\right]+\frac{\sqrt{1-p}}{p}\gamma. \quad (55)$$

It now remains to check whether $\gamma<\lambda^*-\mathbb{E}\left[T\right]$ holds. Using (16), we have

$$\lambda^*-\mathbb{E}\left[T\right]=\frac{\sqrt{1-p}}{p}\gamma-\frac{1-p}{p}\gamma, \quad (56)$$

which is clearly no larger than γ since the quantity $\frac{\sqrt{1-p}-(1-p)}{p}$ is no larger than 1 for all values of p. This gives a contradiction and concludes the proof.

D. Different Methods for Deriving (25)

The method included in the main text to derive (25) involves a calculus of variations approach mainly through leveraging the Euler-Lagrange equation and equating the functional derivative to 0 [32]. In this appendix we discuss two alternate methods to derive (25).

The first method, and quite the simplest one, is by completing the square in the Lagrangian in (23). Specifically, (23) can be rewritten equivalently as

$$\begin{aligned}\mathcal{L} = &\int_c^\gamma \frac{1}{2}\left(w(\tau)+\tau+\mathbb{E}\left[T\right]-\lambda-\frac{\eta(\tau)}{f_Y(\tau)}\right)^2 f_Y(\tau)d\tau \\ &-\int_c^\gamma \frac{1}{2}\left(\tau+\mathbb{E}\left[T\right]-\lambda-\frac{\eta(\tau)}{f_Y(\tau)}\right)^2 f_Y(\tau)d\tau \\ &+\mathbb{E}\left[\overline{Y}\right]\mathbb{E}\left[T\right]+\frac{1}{2}\mathbb{E}\left[T^2\right]-\lambda\mathbb{E}\left[T\right],\end{aligned} \tag{57}$$

which is minimized iff the first integrand is set to 0 $\forall\tau$, which exactly gives (25).

The second method is by using the result in [32, Ch. 7 Th. 1] to conclude that $\mathcal{L}(w)$ is minimized at w^* only if

$$\left.\frac{\partial}{\partial\alpha}\mathcal{L}\left(w^*+\alpha h\right)\right|_{\alpha=0}=0, \tag{58}$$

for any $h(\cdot):\mathbb{R}_+\to\mathbb{R}_+$. Taking $h(\tau)\triangleq\delta(\tau-t)$, for some $t\in[c,\gamma]$, where $\delta(\cdot)$ is the Dirac delta function, we get that for fixed α

$$\begin{aligned}\mathcal{L}(w+\alpha h)=&\int_c^\gamma\left(\left(\tau+\mathbb{E}\left[T\right]-\lambda\right)w(\tau)+\frac{1}{2}w^2(\tau)\right)f_Y(\tau)d\tau \\ &+\alpha\left(t+\mathbb{E}\left[T\right]-\lambda\right)f_Y(t)\int_c^\gamma\delta(\tau-t)d\tau \\ &+\frac{1}{2}\int_c^\gamma\left(w(\tau)+\alpha\delta(\tau-t)\right)^2 f_Y(\tau)d\tau \\ &+\mathbb{E}\left[\overline{Y}\right]\mathbb{E}\left[T\right]+\frac{1}{2}\mathbb{E}\left[T^2\right]-\lambda\mathbb{E}\left[T\right] \\ &-\int_c^\gamma w(\tau)\eta(\tau)d\tau-\alpha\eta(t)\int_c^\gamma\delta(\tau-t)d\tau.\end{aligned} \tag{59}$$

Therefore, upon using $\int_c^\gamma\delta(\tau-t)d\tau=1$, we have

$$\begin{aligned}\frac{\partial\mathcal{L}\left(w+\alpha h\right)}{\partial\alpha}=&\left(t+\mathbb{E}\left[T\right]-\lambda\right)f_Y(t)-\eta(t) \\ &+\int_c^\gamma\left(w(\tau)+\alpha\delta(\tau-t)\right)\delta(\tau-t)f_Y(\tau)d\tau.\end{aligned} \tag{60}$$

Setting $\alpha=0$ in the above and using (58), (25) is directly reached after rearranging.

References

[1] S. K. Kaul, R. D. Yates, and M. Gruteser. Real-time status: How often should one update? In *Proc. IEEE Infocom*, March 2012.

[2] C. Kam, S. Kompella, and A. Ephremides. Age of information under random updates. In *Proc. IEEE ISIT*, July 2013.

[3] 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 Trans. Inf. Theory*, 63(11):7492–7508, November 2017.

[4] R. Talak, S. Karaman, and E. Modiano. Optimizing information freshness in wireless networks under general interference constraints. In *Proc. MobiHoc*, June 2018.

[5] B. Zhou and W. Saad. Optimal sampling and updating for minimizing age of information in the internet of things. In *Proc. IEEE Globecom*, December 2018.

[6] M. Zhang, A. Arafa, J. Huang, and H. V. Poor. How to price fresh data. In *Proc. WiOpt*, June 2019.

[7] M. Bastopcu and S. Ulukus. Minimizing age of information with soft updates. *J. Commun. Netw.*, 21(3):233–243, June 2019.

[8] B. Buyukates, A. Soysal, and S. Ulukus. Age of information in multihop multicast networks. *J. Commun. Netw.*, 21(3):256–267, June 2019.

[9] 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.*, 2(1):193–204, March 2018.

[10] 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*. To appear. Available Online: arXiv:1806.07271.

[11] B. T. Bacinoglu, Y. Sun, E. Uysal-Biyikoglu, and V. Mutlu. Achieving the age-energy tradeoff with a finite-battery energy harvesting source. In *Proc. IEEE ISIT*, June 2018.

[12] T. Z. Ornee and Y. Sun. Sampling for remote estimation through queues: Age of information and beyond. In *Proc. WiOpt*, June 2019.

[13] R. D. Yates, E. Najm, E. Soljanin, and J. Zhong. Timely updates over an erasure channel. In *Proc. IEEE ISIT*, June 2017.

[14] A. Arafa, K. Banawan, K. Seddik, and H. V. Poor. On timely channel coding with hybrid ARQ. In *Proc. IEEE Globecom*, December 2019. Available Online: arXiv:1905.03238.

[15] S. K. Kaul, R. D. Yates, and M. Gruteser. Status updates through queues. In *Proc. CISS*, March 2012.

[16] M. Costa, M. Codreanu, and A. Ephremides. On the age of information in status update systems with packet management. *IEEE Trans. Inf. Theory*, 62(4):1897–1910, April 2016.

[17] K. Chen and L. Huang. Age-of-information in the presence of error. In *Proc. IEEE ISIT*, June 2016.

[18] R. D. Yates and S. K. Kaul. The age of information: Real-time status updating by multiple sources. *IEEE Trans. Inf. Theory*, 65(3):1807–1827, March 2019.

[19] E. Najm and E. Telatar. Status updates in a multi-stream M/G/1/1 preemptive queue. In *Proc. IEEE Infocom*, April 2018.

[20] A. Soysal and S. Ulukus. Age of information in G/G/1/1 systems: Age expressions, bounds, special cases, and optimization. Available Online: arXiv:1905.13743.

[21] S. Farazi, A. G. Klein, and D. R. Brown III. Age of information in energy harvesting status update systems: When to preempt in service? In *Proc. IEEE ISIT*, June 2018.

[22] V. Kavitha abd E. Altman and I. Saha. Controlling packet drops to improve freshness of information. Available Online: arXiv:1807.09325.

[23] B. Wang, S. Feng, and J. Yang. When to preempt? age of information minimization under link capacity constraint. *J. Commun. Netw.*, 2019. To appear.

[24] C. Xu, H. H. Yang, X. Wang, and T. Q. S. Quek. On peak age of information in data preprocessing enabled IoT networks. In *Proc. IEEE WCNC*, April 2019.

[25] Q. Kuang, J. Gong, X. Chen, and X. Ma. Age-of-information for computation-intensive messages in mobile edge computing. Available Online: arXiv:1901.01854.

[26] J. Gong, Q. Kuang, X. Chen, and X. Ma. Reducing age-of-information for computation-intensive messages via packet replacement. Available Online: arXiv:1901.04654.

[27] P. Zou, O. Ozel, and S. Subramaniam. Trading off computation with transmission in status update systems. Available Online: arXiv:1907.00928.

[28] X. Song, X. Qin, Y. Tao, B. Liu, and P. Zhang. Age based task scheduling and computation offloading in mobile-edge computing systems. Available Online: arXiv:1905.11570.

[29] R. D. Yates, M. Tavan, Y. Hu, and D. Raychaudhuri. Timely cloud gaming. In *Proc. IEEE Infocom*, May 2017.

[30] W. Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967.

[31] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[32] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.