Age-Delay Tradeoffs in Single Server Systems

Rajat Talak and Eytan Modiano

***Abstract*—Information freshness and low latency communication is important to many emerging applications. While Age of Information (AoI) serves as a metric of information freshness, packet delay is a traditional metric of communication latency. We prove that there is a natural tradeoff between the AoI and packet delay. We consider a single server system, in which at most one update packet can be serviced at a time. The system designer controls the order in which the packets get serviced and the service time distribution, with a given service rate. We analyze two tradeoff problems that minimize packet delay and the variance in packet delay, respectively, subject to an average age constraint. We prove a strong age-delay and age-delay variance tradeoff, wherein, as the average age approaches its minimum, the delay and its variance approach infinity. We show that the service time distribution that minimizes average age, must necessarily have an unbounded-second moment.**

I. Introduction

Information freshness and low latency communication is gaining increasing relevance in many communication systems [1]. Age of information (AoI) is a newly proposed metric of information freshness [2], while packet delay is a traditional metric of latency in communication. AoI measures the time since the last received fresh update was generated at the source, and is therefore a destination centric metric. It only accounts for packets that deliver fresh updates to the destination. Packet delay, unlike AoI, is a packet centric metric that takes into account the delay incurred by each packet in the system. In this work, we show that there is a natural tradeoff between the two metrics of AoI and packet delay.

AoI was first studied for the first come first serve (FCFS) M/M/1, M/D/1, and D/M/1 queues in [2]. Since then, AoI has been analyzed for several queueing systems [2]–[13], with the goal to minimize AoI. Two time average metrics of AoI, namely, peak and average age are generally considered.

The advantage of having parallel servers, towards improving AoI, was demonstrated in [3], [4], [9]. Having smaller buffer sizes [10], [11] or introducing packet deadlines [11]–[13], in which a packet deletes itself after its deadline expiration, are two other considered ways of improving AoI. In [6], the LCFS queue scheduling discipline, with preemptive service, is shown to be an age optimal, when the service times are exponentially distributed. AoI for the LCFS queue with Poisson arrivals and Gamma distributed service was analyzed in [5].

In most of these works, optimal update generation and service rate is sought that minimizes age, and in several, the question of determining a good queue scheduling discipline for AoI minimization is considered with interest. More recently, determining optimal update generation and service time distribution, that minimizes AoI, has been studied in [14]. In [14], [15], we showed that for the LCFS queue with preemptive service (LCFSp) and G/G/∞ queue, a heavy tailed service time distribution minimizes AoI. It is noteworthy that such a heavy tailed service maximizes packet delay for the LCFSp queue and packet delay variance for the G/G/∞ queue, respectively.

This points to a natural tradeoff between age and delay. In Figure 1, we plot the achieved age-delay point under various queue scheduling disciplines and service time distributions. It appears that lower age can be achieved but only at a cost of higher delay. In this work, we prove that there is, in fact, a tradeoff between age and delay.

We consider a single server system, that can service at most one packet at any given time. Generated updates are sent to this single server system. We assume that the system designer decides the queue scheduling discipline, i.e. the order in which the packets get serviced, and the service time distribution. Note that the service time distribution generally depends on the packet length, and therefore, a given packet length distribution

The authors are with the Laboratory for Information and Decision Systems (LIDS) at the Massachusetts Institute of Technology (MIT), Cambridge, MA. {talak, modiano}@mit.edu

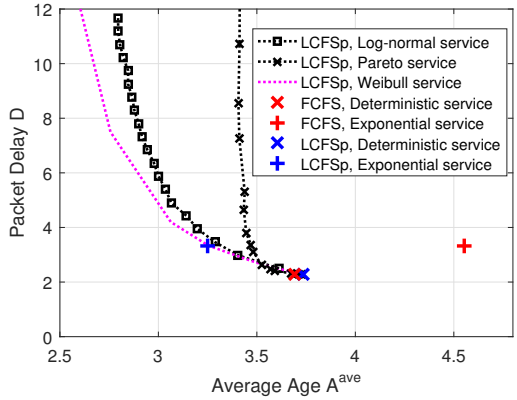

Fig. 1. Plot of achieved age-delay points for various single server systems, Poisson packet generation at rate $\lambda = 0.5$, and service at rate $\mu = 0.8$. Scheduling disciplines: FCFS, LCFSp. Service time distributions: Deterministic, Exponential, and Heavy Tailed distributions in Table I.

TABLE I
Heavy tailed service time distributions with mean $\mathbb{E}[S] = 1/\mu$.

Name	Distribution	Free Parameter
Log-normal	$S = \exp\left(-\log\mu - \frac{\sigma^2}{2} + \sigma N\right)$ $N \sim \mathcal{N}(0,1)$	$\sigma > 0$
Pareto	$F_S(x) = 1 - (\theta(\alpha)/x)^{\alpha}\mathbb{I}_{\{x>\theta(\alpha)\}}$ $\theta(\alpha) = (\alpha-1)/(\mu\alpha)$	$\alpha > 1$
Weibull	$\mathbf{P}[S > x] = \exp\{-(x/\beta(k))^k\}$ $\beta(k) = [\mu\Gamma(1+1/k)]^{-1}$	$k > 0$

can be induced on the generated update packets.

We consider the problem of minimizing packet delay, subject to an average age constraint, over the space of all queue scheduling disciplines and service time distributions, with a fixed mean service time budget of $1/\mu$. For a given update generation process, we show that there is a strong age-delay tradeoff, namely, as the average age approaches its minimum, the delay approaches infinity. The same result holds also for packet delay variance, i.e. as the average age approaches its minimum, the variance in packet delay approaches infinity. We also consider two restrictions on the system model, for which the age-delay tradeoff vanishes.

The system model and the age-delay tradeoff problems are described in Section II. Strong age-delay and age-delay variance tradeoff is proved in III. In Section IV, we consider the two specific instances when the age-delay tradeoff vanishes, and conclude in Section V.

II. System Model

A source generates update packets according to a renewal process, at a given rate λ. The update packets enter a queueing system. The server has rate μ, and can service at most one update packet at any given time. The service times are independent and identically distributed across update packets.

The system designer has control over two things: It can decide the service time distribution, and it can decide the order in which the update packets get serviced. We assume that in determining the order of service, the scheduler is not privy to the service times of the individual packets. The scheduler is also not allowed to drop any packets. Since we are concerned about age and packet delay, we will assume that the arrival rate is less than the service rate: $\lambda < \mu$.

We use X to denote the inter-generation time of update packets with distribution F_X, and S to denote the service time random variable, with distribution F_S. We use Minimize or $\min$, instead of the technically correct $\inf$, for ease of presentation. We now define the two latency metrics of average age of information and packet delay.

A. Delay and Age of Information

Let the update packets be generated at times $t_1, t_2, \ldots$, and let the update packet i reach the destination at time $t_i^{'}$. The update packets may not reach the destination in the same order as they were generated. In Figure 2, packet 3 reaches the destination before packet 2, i.e. $t_3^{'} < t_2^{'}$. Delay for the ith packet is $D_i = t_i^{'} - t_i$, and the packet delay for the system is given by

$$D = \limsup_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} D_i\right], \tag{1}$$

where the expectation is over the update generation, service times, and scheduling discipline. We skip a formal definition, but will use the notation VarD to denote variance in packet delay. For a single server queueing system, we note that VarD is lower-bounded by variance in service time distribution S.

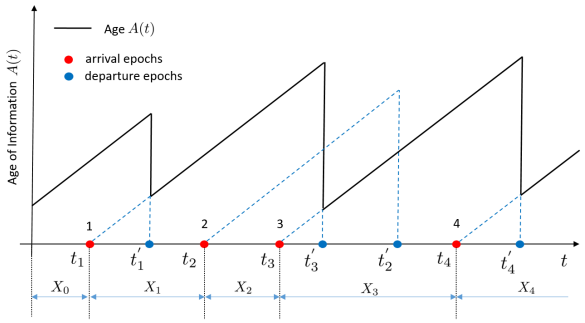

Fig. 2. Age evolution in time. Here, t_i and $t_i^{'}$ denote the generation and reception times of packet i.

Age of a packet i is defined as the time since it was generated: $A^i(t) = (t - t_i)\mathbb{I}_{\{t>t_i\}}$, which is 0 by definition for time prior to its generation $t < t_i$. Age of information at the destination node, at time t, is defined as the minimum age across all received packets up to time t:

$$A(t) = \min_{i\in\mathcal{P}(t)} A^i(t), \tag{2}$$

where $\mathcal{P}(t) \subset \{1, 2, 3, \ldots\}$ denotes the set of packets received by the destination, up to time t. Notice that AoI increases linearly, and drops only at the times of certain packet receptions: $t_1^{'}, t_3^{'}, t_4^{'}, \ldots$, but not $t_2^{'}$ in Figure 2. Such an age drop occurs only when an update packet with a lower age, than all packets received thus far, is received by the destination. We refer to such packets, that result in age drops, as the *informative packets* [3]. The average age is defined to be the time averaged area under the age curve:

$$A^{\text{ave}} = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T A(t)dt\right], \tag{3}$$

where the expectation is over the packet generation and packet service processes.

We use the notation $D(F_S, \pi_Q)$, $\text{VarD}(F_S, \pi_Q)$, and $A^{\text{ave}}(F_S, \pi_Q)$ to make explicit the dependency of delay, its variance, and average age, respectively, on the service time distribution F_S and the queue scheduling policy π_Q.

B. Age-Delay Tradeoff Problems

We define two age-delay tradeoff problems. One, minimizes delay while the other minimizes delay variance, both over an average age constraint. The age-delay tradeoff is defined as:

$$\begin{aligned} T(\text{AoI}) = \underset{F_S,\pi_Q}{\text{Minimize}} \quad & D(F_S, \pi_Q) \\ \text{subject to} \quad & A^{\text{ave}}(F_S, \pi_Q) \le \text{AoI}, \\ & \mathbb{E}[S] = 1/\mu. \end{aligned} \tag{4}$$

Here, the function $T(\text{AoI})$ denotes the minimum delay that can be achieved for the single server queueing system, with an average age constraint of $A^{\text{ave}}(F_S, \pi_Q) \le \text{AoI}$. It might seem that both minimum age and delay could be attained

simultaneously. We will show that, $T(\text{AoI}) \to \infty$ as AoI approaches the minimum average age

$$A_{\min} = \underset{F_S,\pi_Q}{\text{Minimize}} \quad A^{\text{ave}}(F_S,\pi_Q). \tag{5}$$

In [14], we proved such a result for the LCFS queues with preemptive service (LCFSp). In this work, we show that such a result holds, even when the system designer has an option of choosing the queue scheduling discipline π_Q. This does not follow trivially from the LCFSp result in [14], primarily because LCFSp is not known to be the optimal scheduling discipline for single server systems, especially when the service times are not exponentially distributed [6], [16].

Variability in packet delay is also an important metric in system performance. We define the age-delay variance tradeoff problem to be:

$$\begin{aligned} V(\text{AoI}) = \underset{F_S,\pi_Q}{\text{Minimize}} \quad & \text{VarD}(F_S,\pi_Q) \\ \text{subject to} \quad & A^{\text{ave}}(F_S,\pi_Q) \le \text{AoI}, \\ & \mathbb{E}[S] = 1/\mu. \end{aligned} \tag{6}$$

Here, the function $V(\text{AoI})$ denotes the minimum delay variance that can be achieved for the single server queueing system, with an average age constraint of $A^{\text{ave}}(F_S,\pi_Q) \le \text{AoI}$. Counter to our intuition, we show that $V(\text{AoI}) \to +\infty$ as AoI approaches its minimum value $A_{\min}$.

III. Age-Delay Tradeoff

Ideally, we would like to obtain every point on the tradeoff curves, i.e., completely characterize the functions: $T(\text{AoI})$ and $V(\text{AoI})$. The following result motivates optimization of a linear combination of average age and packet delay, in order to achieve every point on the tradeoff curve.

Theorem *1:* The points on the age-delay tradeoff curve $T(\text{AoI})$ can be obtained by solving

$$\begin{aligned} \underset{F_S,\pi_Q}{\text{Minimize}} \quad & D(F_S,\pi_Q) + \nu A^{\text{ave}}(F_S,\pi_Q) \\ \text{subject to} \quad & \mathbb{E}[S] = 1/\mu, \end{aligned} \tag{7}$$

for all $\nu \ge 0$. Similarly, the points on the age-delay variance tradeoff curve $V(\text{AoI})$ can be obtained by solving (7), by replacing $D(F_S,\pi_Q)$ with $\text{VarD}(F_S,\pi_Q)$.

Proof: The proof uses simple duality arguments, and is omitted due to space constraints. ∎

Theorem 1 motivates optimization of a latency metric that is a linear combination of average age and packet delay (or packet delay variance). This problem, however, is not easy to solve for the following reason: the delay is minimized with deterministic service times, while the variance in delay is minimized under FCFS service discipline [17]. The opposite holds for the average age: LCFSp queue scheduling policy with heavy tailed service distribution is known to achieve minimum age [14], [15]. Thus, the delay term and the average age term in (7) pull the decision variables in opposite directions. In what follows, we prove that there is a strong age-delay tradeoff.

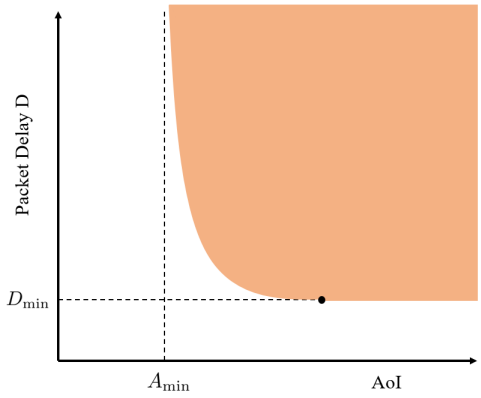

Fig. 3. Illustration of strong age-delay tradeoff.

We say that a *strong age-delay tradeoff* exists for $T(\text{AoI})$ if $T(\text{AoI}) \to +\infty$ and AoI approaches $A_{\min}$. Conversely, *no age-delay tradeoff* exists for $T(\text{AoI})$ if the minimum average age and the minimum packet delay can be achieved simultaneously. Similar definition apply for age-delay variance tradeoff $V(\text{AoI})$.

Figure 3 illustrates a strong age-delay tradeoff. Note that this matches with our numerical results in Figure 1. In what follows, we prove a strong tradeoff between age-delay and age-delay variance. We first derive the minimum average age $A_{\min}$, over the space of all scheduling policies and service time distributions.

Lemma *1:* The minimum average age $A_{\min} = \frac{1}{2}\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]}$.

Proof: The fact that $\frac{1}{2}\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]}$ is a lower-bound on the average age, can be proved by pretending that each update packet spends zero time in the system, i.e. $t_i = t_i'$. This provides a sample path lower bound for the age process. In this sample path, the age drops to 0 at every t_i, and increases to $t_{i+1} - t_i$, just before dropping to 0 again at t_{i+1}. The average age of this artificially constructed, lower-bound age process is $\frac{1}{2}\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]}$, which implies $A_{\min} \ge \frac{1}{2}\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]}$.

In [14], we proved that this age lower-bound is achieved by the LCFSp queue scheduling policy and heavy tailed service time distributions in Table I: the Pareto, log-normal, and Weibull distributed service attain the lower-bound as $\alpha \downarrow 1$, $\sigma \uparrow +\infty$, and $k \downarrow 0$, respectively [14]. ∎

We now prove the strong age-delay tradeoff and age-delay variance tradeoff.

Theorem *2:* There is a strong age-delay tradeoff and age-delay variance tradeoff, namely, $T(\text{AoI}) \to +\infty$ and $V(\text{AoI}) \to +\infty$ as $\text{AoI} \to A_{\min}$.

Proof: **1. Age-delay tradeoff:** First, note that the packet delay is given by $D(F_S, \pi_Q) = \frac{\lambda}{2}\frac{\mathbb{E}\left[S^2\right]}{1-\rho} + \mathbb{E}\left[S\right]$, where $\rho = \frac{\lambda}{\mu}$, for any scheduling policy π_Q that does not use the individual packet service times to schedule it. It, therefore, suffices to show that we have $\mathbb{E}\left[S^2\right] \to +\infty$ as AoI $\to A_{\min}$.

We first note that the average age $A^{\text{ave}}(F_S, \pi_Q)$, under any queue scheduling policy π_Q, is lower-bounded by the average age for the G/G/∞ queue:

$$A^{\text{ave}}(F_S, \pi_Q) \geq A^{\text{ave}}_{\text{G/G/}\infty}. \tag{8}$$

This is because, in G/G/∞ queue, an arriving packet is immediately put to service, and therefore, incurs no queueing delay. Due to this the average age for the G/G/∞ queue serves as a lower-bound for any single server queue, in a stochastic ordering sense. Taking expected value yields (8).

From [14], we know the average age for the G/G/∞ queue to be:

$$A^{\text{ave}}_{\text{G/G/}\infty} = \frac{1}{2}\frac{\mathbb{E}\left[X^2\right]}{\mathbb{E}\left[X\right]} + \mathbb{E}\left[\min_{l\geq 0}\left(\sum_{k=1}^{l} X_k + S_{l+1}\right)\right]. \tag{9}$$

Notice that the first term in (9) is nothing but $A_{\min}$. Therefore, as AoI $\to$ $A_{\min}$ in (4), it must be the case that $\mathbb{E}\left[\min_{l\geq 0}\left(\sum_{k=1}^{l} X_k + S_{l+1}\right)\right] \to 0$. Lemmas 2 and 3, in Appendix A, prove that $\mathbb{E}\left[\min_{l\geq 0}\left(\sum_{k=1}^{l} X_k + S_{l+1}\right)\right] \to 0$ implies $\mathbb{E}\left[S^2\right] \to +\infty$.

2. Age-delay variance tradeoff: Variance VarD(F_S, π_Q) is lower-bounded by the variance in service time $\mathbb{E}\left[S^2\right] - \mathbb{E}\left[S\right]^2$ for any queue scheduling policy π_Q. It therefore suffices to argue that as AoI $\to A_{\min}$ we have $\mathbb{E}\left[S^2\right] \to +\infty$, which we just proved to be true. ∎

In the proof, we essentially showed that $\mathbb{E}\left[S^2\right] \to +\infty$ is a necessary condition for the average age to approach the minimum $A_{\min}$. It seems counterintuitive at first that a strong tradeoff should exist between delay, or delay variance, and average age. However, a close examination reveals the following insight:

For age minimization it becomes necessary that the informative packets, the packets that reduce age, get serviced as soon as they arrive, while the non-informative packets, may incur as long a service time and queueing delay, as they do not matter in the age calculations. As we do this, the packet delay gets dominated by the delay of the non-informative packets, resulting in the two age-delay tradeoffs.

We have assumed that the packet inter-generation time distribution to be fixed. The results in Lemma 1 and Theorem 2 imply that a strong age-delay tradeoffs will hold even if we could control the inter-generation time distribution F_X, with a mean budget of $\mathbb{E}\left[X\right] = 1/\lambda$. The optimal F_X would be deterministic as $A_{\min} = \frac{1}{2}\frac{\mathbb{E}\left[X^2\right]}{\mathbb{E}\left[X\right]^2} \geq \frac{1}{2}\mathbb{E}\left[X\right]$, thus, making periodic generation of updates optimal.

IV. Special Cases of No Tradeoff

In the previous section, we proved a strong age-delay and age-delay variance tradeoff. We now consider two scenarios of the single server system, for which the age-delay tradeoff vanishes, i.e. the minimum age and minimum delay can be attained simultaneously.

1) Memoryless Service Times: Consider a system in which service times are exponentially distributed. The system designer has to decide only the queue scheduling policy π_Q that solves (4). We know from the works in [6], [16] that LCFSp minimizes average age, when the service times are exponentially distributed. The queueing delay $D(F_S, \pi_Q)$ remains the same, under any queue scheduling policy π_Q, that does not use individual packet service times to schedule [18]. Thus, the minimum age and minimum delay is achieved simultaneously. A version of this result was proved in [19].

> **Theorem** *3 ([19]):* If service times are exponentially distributed, then there is no age-delay tradeoff.

2) FCFS Queue Schedule: FCFS queue scheduling is used in many practical systems [10], [20]. Periodic update generation is also known to reduces age in these systems. Consider the case of periodic update generation and FCFS queue scheduling.

Deterministic service is known to minimize packet delay for the FCFS queue scheduling [18]. It is also known that periodic generation and deterministic service minimized average age for the FCFS queue [7], [14]: $A^{\text{ave}}_{\text{D/D/1}} \leq A^{\text{ave}}_{\text{G/G/1}}$. This gives us the following result.

> **Theorem** *4:* If the update generation is periodic and queue scheduling policy is FCFS, then there is no age-delay tradeoff.

V. Conclusion

We considered a single server system, in which at most a single packet can be serviced at any given time. The system designer decides the order in which arriving packets get serviced and the service time distribution, with a given mean service time budget. We proved a strong age-delay and age-delay variance tradeoff, wherein as age approaches its minimum, the delay and its variance approach infinity.

We note the following reason for the tradeoff: For age optimality, informative packets, which reduce age, need to be serviced quickly, whereas a long service time and queuing delay can be incurred by other non-informative packets. As we do this, the packet delay, and its variance, get dominated by the delay of the non-informative packets. This leads to the age-delay tradeoff.

References

[1] X. Jiang, H. Shokri-Ghadikolaei, G. Fodor, E. Modiano, Z. Pang, M. Zorzi, and C. Fischione, "Low-latency networking: Where latency lurks and how to tame it," *Proc. of the IEEE*, pp. 1–27, Aug. 2018.

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

[3] C. Kam, S. Kompella, and A. Ephremides, "Effect of message transmission diversity on status age," in *Proc. ISIT*, pp. 2411–2415, Jun. 2014.

[4] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," in *Proc. ISIT*, pp. 1583–1587, Jun. 2014.
[5] E. Najm and R. Nasser, "Age of information: The gamma awakening," *ArXiv e-prints*, Apr. 2016.
[6] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Minimizing the age of the information through queues," *arXiv e-prints arXiv:1709.04956*, Sep. 2017.
[7] Y. Inoue, H. Masuyama, T. Takine, and T. Tanaka, "The stationary distribution of the age of information in FCFS single-server queues," in *Proc. ISIT*, pp. 571–575, Jun. 2017.
[8] A. Soysal and S. Ulukus, "Age of information in G/G/1/1 systems," *arXiv e-prints arXiv:1805.12586*, Jun. 2018.
[9] R. D. Yates, "Status updates through networks of parallel servers," in *Proc. ISIT*, pp. 2281–2285, Jun. 2018.
[10] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *Proc. SECON*, pp. 350–358, Jun. 2011.
[11] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Controlling the age of information: Buffer size, deadline, and packet replacement," in *Proc. MILCOM*, pp. 301–306, Nov. 2016.
[12] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Age of information with a packet deadline," in *Proc. ISIT*, pp. 2564–2568, Jul. 2016.
[13] Y. Inoue, "Analysis of the age of information with packet deadline and infinite buffer capacity," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 2639–2643, Jun. 2018.
[14] R. Talak, S. Karaman, and E. Modiano, "Can determinacy minimize age of information?," *arXiv e-prints arXiv:1810.04371*, Oct. 2018.
[15] R. Talak, S. Karaman, and E. Modiano, "When a heavy tailed service minimizes age of information," in *Submitted to ISIT*, Jul. 2019.
[16] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Age-optimal information updates in multihop networks," in *Proc. ISIT*, pp. 576–580, Jun. 2017.
[17] J. F. C. Kingman, "The effect of queue discipline on waiting time variance," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58, no. 1, p. 163164, 1962.
[18] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Prentice Hall, 2 ed., 1992.
[19] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Optimizing data freshness, throughput, and delay in multi-server information-update systems," in *Proc. ISIT*, pp. 2569–2573, Jul. 2016.
[20] R. Talak, S. Karaman, and E. Modiano, "Optimizing information freshness in wireless networks under general interference constraints," in *Proc. Mobihoc*, Jun. 2018.

APPENDIX

A. Properties of Service Time Random Variable S

Let S be a continuous random variable with distribution F_S, with parameter η, such that $\mathbb{E}[S] = 1/\mu$ for all η. We would like to derive conditions on S such that

$$\mathbb{E}\left[\min_{l\geq 0}\left(\sum_{k=1}^{l} X_k + S_{l+1}\right)\right] \to 0,$$

as η approaches certain η^*, for a given distribution F_X. We now derive an equivalent condition that only requires verifying certain properties of F_S.

Lemma *2:* For S_l and X_k that are i.i.d. distributed according to F_S and F_X, respectively, we have

$$\lim_{\eta\to\eta^*} \mathbb{E}\left[\min_{l\geq 0}\left(\sum_{k=1}^{l} X_k + S_{l+1}\right)\right] = 0, \qquad (10)$$

if and only if, for all $x \geq 1/\lambda$, we have

$$\lim_{\eta\to\eta^*} \mathbf{P}[S > x] = 0, \text{ and } \lim_{\eta\to\eta^*} \mathbb{E}\left[S\mathbb{I}_{\{S<x\}}\right] = 0. \quad (11)$$

Proof: The fact that (11) imply (10) is proved in our recent work [14], [15]. Here, we establish that (10) implies the conditions (11) on distribution F_S.

Let $Z = \min_{l\geq 0}\left(\sum_{k=1}^{l} X_k + S_{l+1}\right)$. We first lower-bound Z as follows:

$$Z = \min\{S_1, X_1+S_2, X_1+X_2+S_3, \ldots\} = \min\{S_1, X_1+Z'\},$$

where $Z' = \min\{S_2, X_2+S_3, X_2+X_3+S_4, \ldots\}$. Since $Z' \geq 0$, we must have $Z \geq \min\{S_1, X_1\}$. Without loss of generality, we can lose the subscripts and write $Z \geq \min\{S, X\}$, where $S \sim F_S$ and $X \sim F_X$.

If $\mathbb{E}[Z] \to 0$ as $\eta \to \eta^*$ then clearly $\mathbb{E}[\min\{S, X\}] \to 0$ as $\eta \to \eta^*$. Now construct $\hat{X}$ such that:

$$\hat{X} = \begin{cases} 0 & \text{if } X < 1/\lambda \\ 1/\lambda & \text{if } X \geq 1/\lambda \end{cases}.$$

Clearly, $\hat{X} \leq X$, and thus, $\min\{S, \hat{X}\} \leq \min\{S, X\}$, which implies $\mathbb{E}\left[\min\{S, \hat{X}\}\right] \to 0$. Since $\hat{X}$ takes only two values, namely 0 and $1/\lambda$, we have $\mathbb{E}\left[\min\{S, \hat{X}\}\right] = \mathbb{E}[\min\{S, 1/\lambda\}]\,\mathbf{P}[X \geq 1/\lambda]$. Now, $\mathbf{P}[X \geq 1/\lambda] > 0$ because $\mathbb{E}[X] = 1/\lambda$. Further, $\mathbf{P}[X \geq 1/\lambda]$ does not depend on S, and therefore, is also independent of the parameter η. Therefore, $\mathbb{E}\left[\min\{S, \hat{X}\}\right] \to 0$ implies $\mathbb{E}[\min\{S, 1/\lambda\}] \to 0$. Using monotonicity of $\mathbf{P}[S > x]$ in x one can show that $\mathbb{E}[\min\{S, 1/\lambda\}] \to 0$ implies

$$\lim_{\eta\to\eta^*} \mathbb{E}[\min\{S, x\}] = 0, \qquad (12)$$

for all $x \geq 1/\lambda$. Now, notice that

$$\mathbb{E}[\min\{S, x\}] = \mathbb{E}\left[S\mathbb{I}_{\{S<x\}}\right] + x\mathbb{E}\left[\mathbb{I}_{\{S>x\}}\right], \qquad (13)$$

where we can ignore the equality case $S = x$ since S is continuously distributed. Substituting (13) in (12) we obtain (11). ∎

We now give a sufficient condition on the service time distributions F_S, parameterized by η, to have its second moment approach infinity.

Lemma *3:* For the parameterized, service time random variable S, we have $\lim_{\eta\to\eta^*} \mathbb{E}\left[S^2\right] = +\infty$ if

$$\lim_{\eta\to\eta^*} \mathbf{P}[S > x] = 0, \text{ and } \lim_{\eta\to\eta^*} \mathbb{E}\left[S\mathbb{I}_{\{S<x\}}\right] = 0, \quad (14)$$

for all $x > x_0$, and some $x_0 > 0$.

Proof: Let the two conditions (14) hold for S. First, note that $\mathbb{E}\left[S^2\right] \geq \mathbb{E}\left[S^2\mathbb{I}_{\{S>x\}}\right] \geq x\mathbb{E}\left[S\mathbb{I}_{\{S>x\}}\right]$, for all $x > 0$. We can write $\mathbb{E}\left[S\mathbb{I}_{\{S>x\}}\right]$ as $\mathbb{E}[S] - \mathbb{E}\left[S\mathbb{I}_{\{S<x\}}\right] \to 1/\mu$ as $\eta \to \eta^*$ by (14) and the fact that $\mathbb{E}[S] = 1/\mu$. Therefore, we have $\liminf_{\eta\to\eta^*} \mathbb{E}\left[S^2\right] \geq x/\mu$ for all $x > x_0$. This can only be true if $\lim_{\eta\to\eta^*} \mathbb{E}\left[S^2\right] = +\infty$. ∎