Age-Optimal Updates of Multiple Information Flows

Yin Sun[†], Elif Uysal-Biyikoglu[‡], and Sastry Kompella[*]
[†]Dept. of ECE, Auburn University, Auburn, AL
[‡]Dept. of EEE, Middle East Technical University, Ankara, Turkey
[*]Information Technology Division, Naval Research Laboratory, Washington, DC

***Abstract*—In this paper, we study an age of information minimization problem, where multiple flows of update packets are sent over multiple servers to their destinations. Two online scheduling policies are proposed. When the packet generation and arrival times are synchronized across the flows, the proposed policies are shown to be (near) optimal for minimizing any *time-dependent*, *symmetric*, and *non-decreasing* penalty function of the ages of the flows over time in a stochastic ordering sense.**

I. Introduction

In many information-update and networked control systems, such as news updates, stock trading, autonomous driving, and robotics control, information has the greatest value when it is fresh. A metric on information freshness, called the *age of information* or simply the *age*, was defined in [1], [2]. Consider a flow of update packets that are sent from a source to a destination through a queue. Let $U(t)$ be the time stamp (i.e., generation time) of the newest update that the destination has received by time t. The age of information, as a function of time t, is defined as $\Delta(t) = t - U(t)$, which is the time elapsed since the newest update was generated.

In recent years, there have been a lot of research efforts on the behavior of $\Delta(t)$ under various queueing service disciplines and how to control $\Delta(t)$ to keep the information as fresh as possible [2]–[15]. When there is a single flow of update packets, a Last Generated First Served (LGFS) update transmission policy, in which the last generated packet is served the first, has been shown to be (nearly) optimal for minimizing the age process $\{\Delta(t), t \geq 0\}$ in a stochastic ordering sense for multi-server and multi-hop networks [5]–[8]. This result holds for arbitrary packet generation times at the source and arbitrary packet arrival times at the transmitter queue (e.g., the queue in Fig. 1); it also holds for minimizing any non-decreasing functional $p(\{\Delta(t), t \geq 0\})$ of the age process. These studies motivated us to explore service and scheduling policies for achieving age optimality in more general systems with *multiple flows of update packets*. In this case, the transmission scheduler needs to compare not only the packets from the same flow, but also the packets from different flows, which makes the scheduling problem more challenging.

In this paper, we study age-optimal online scheduling in multi-flow, multi-server queueing systems (as illustrated in Figure 1), where each server can be used to send update packets to any destination, one packet at a time. We assume that the packet generation and arrival times are synchronized across the flows. This assumption is a generalized version of the model in [12]. In practice, synchronized update generations and arrivals occur when there is a single source and multiple destinations (e.g., [12]), or in periodic sampling where multiple sources are synchronized by the same clock as in many monitoring and control applications. The contributions of this paper are summarized as follows:

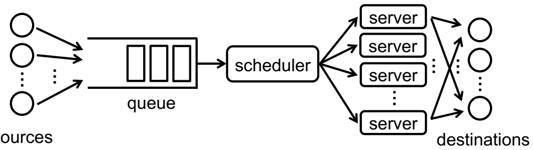

Fig. 1. System model.

- Let $\boldsymbol{\Delta}(t)$ denote the age vector of multiple flows. We introduce an age penalty function $p_t(\boldsymbol{\Delta}(t))$ to represent the level of dissatisfaction for having aged information at the destinations at time t, where p_t can be any *time-dependent*, *symmetric*, and *non-decreasing* function of the age vector $\boldsymbol{\Delta}(t)$.
- For single-server systems with *i.i.d.* exponential service times, we propose a *Maximum Age First, Last Generated First Served (MAF-LGFS) policy*. If the packet generation and arrival times are synchronized across the flows, then for all age penalty functions p_t defined above, the preemptive MAF-LGFS policy is proven to minimize the age penalty process $\{p_t(\boldsymbol{\Delta}(t)), t \geq 0\}$ among all causal policies in a stochastic ordering sense (Theorem 1).
- For multi-server systems with *i.i.d.* New-Better-than-Used (NBU) service times (which include exponential service times as a special case), we consider an age lower bound called the *Age of Served Information* and propose a *Maximum Age of Served Information First, Last Generated First Served (MASIF-LGFS) policy*. For synchronized packet generations and arrivals, the non-preemptive MASIF-LGFS policy is shown to be within an additive gap from the optimum for minimizing the long-run average age of the flows, where the gap is equal to the mean service time of one packet (Theorems 2-3). Numerical evaluations are provided to verify our (near) age optimality results. Some possible extensions are discussed at the end of the paper.

II. Related Work

The age performance of multiple sources has been analyzed in [9]–[11]. In [15], status updates over a multiaccess channel was studied. In [14], an age minimization problem for single-hop wireless networks with interference constraints was shown to be NP hard, and tractable cases were identified. In [12], the expected time-average of the weighted sum age of multiple sources was minimized in a broadcast network with an ON-OFF channel and periodic arrivals, where only one source is scheduled at a time and the scheduler does not know the current ON-OFF channel state. When the network is symmetric and the weights are equal, a sample-path method was used to show that the maximum age first (MAF) policy is optimal.

Y. Sun's work is supported in part by ONR grant N00014-17-1-2417. S. Kompella's work is supported in part by ONR.

Further, a sub-optimal Whittle's index method was used to handle the general asymmetric cases. In [13], for symmetric Bernoulli arrivals and an always-ON channel with no buffers, the MAF policy was shown to be optimal for minimizing the expected time-average of the sum age of multiple sources. In addition, Markov decision process (MDP) methods were used to handle the general scenarios with asymmetric arrivals and a buffer, where the optimal policies are shown to be switch-type.

Compared with these prior studies, Theorem 1 in this paper may be seen as an extension of the optimal scheduling results in [12], [13] to general time-dependent, symmetric, and non-decreasing age penalty functions p_t. In Theorems 2-3, we go one step forward to study multi-flow, multi-server scheduling, which was not considered in [12], [13]. This paper also complements the studies in [5]–[8] on (near) age-optimal online scheduling with a single information flow.

III. System Model

A. Notation and Definitions

We use lower case letters such as x and $\boldsymbol{x}$, respectively, to represent deterministic scalars and vectors. In the vector case, a subscript will index the components of a vector, such as x_i. We use $x_{[i]}$ to denote the i-th largest component of vector $\boldsymbol{x}$. Let $\mathbf{0}$ denote the vector with all 0 components. A function $f : \mathbb{R}^n \to \mathbb{R}$ is termed *symmetric* if $f(\boldsymbol{x}) = f(x_{[1]}, \ldots, x_{[n]})$ for all $\boldsymbol{x}$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is termed *separable* if there exists functions $f_1, \ldots, f_n$ of one variable such that $f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(x_i)$. The composition of functions f and g is denoted by $f \circ g(\boldsymbol{x}) = f(g(\boldsymbol{x}))$. For any n-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, the elementwise vector ordering $x_i \leq y_i$, $i = 1, \ldots, n$, is denoted by $\boldsymbol{x} \leq \boldsymbol{y}$. Let $\mathcal{A}$ and $\mathcal{U}$ denote sets and events. For all random variable X and event $\mathcal{A}$, let $[X|\mathcal{A}]$ denote a random variable with the conditional distribution of X for given $\mathcal{A}$.

Definition 1. *Stochastic Ordering of Random Variables [16]*: A random variable X is said to be *stochastically smaller* than another random variable Y, denoted by $X \leq_{\text{st}} Y$, if

$$\Pr(X > t) \leq \Pr(Y > t), \ \forall \ t \in \mathbb{R}.$$

Definition 2. *Stochastic Ordering of Random Vectors [16]*: A set $\mathcal{U} \subseteq \mathbb{R}^n$ is called *upper*, if $\boldsymbol{y} \in \mathcal{U}$ whenever $\boldsymbol{y} \geq \boldsymbol{x}$ and $\boldsymbol{x} \in \mathcal{U}$. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two n-dimensional random vectors, $\boldsymbol{X}$ is said to be *stochastically smaller* than $\boldsymbol{Y}$, denoted by $\boldsymbol{X} \leq_{\text{st}} \boldsymbol{Y}$, if

$$\Pr(\boldsymbol{X} \in \mathcal{U}) \leq \Pr(\boldsymbol{Y} \in \mathcal{U}), \ \forall \ \mathcal{U} \subseteq \mathbb{R}^n.$$

Definition 3. *Stochastic Ordering of Stochastic Processes [16]*: Let $\{X(t), t \in [0, \infty)\}$ and $\{Y(t), t \in [0, \infty)\}$ be two stochastic processes, $\{X(t), t \in [0, \infty)\}$ is said to be *stochastically smaller* than $\{Y(t), t \in [0, \infty)\}$, denoted by $\{X(t), t \in [0, \infty)\} \leq_{\text{st}} \{Y(t), t \in [0, \infty)\}$, if for all integer n and $0 \leq t_1 < t_2 < \ldots < t_n$, it holds that

$$(X(t_1), X(t_2), \ldots, X(t_n)) \leq_{\text{st}} (Y(t_1), Y(t_2), \ldots, Y(t_n)).$$

Let $\mathbb{V}$ be the set of Lebesgue measurable functions on $[0, \infty)$, i.e.,

$$\mathbb{V} = \{f : [0, \infty) \mapsto \mathbb{R} \text{ is Lebesgue measurable}\}. \quad (1)$$

A functional $\phi : \mathbb{V} \mapsto \mathbb{R}$ is said to be *non-decreasing* if $\phi(f_1) \leq \phi(f_2)$ for all $f_1, f_2 \in \mathbb{V}$ satisfying $f_1(t) \leq f_2(t)$ for $t \in [0, \infty)$. We remark that $\{X(t), t \in [0, \infty)\} \leq_{\text{st}} \{Y(t), t \in [0, \infty)\}$ if, and only if, [16]

$$\mathbb{E}[\phi(\{X(t), t \in [0, \infty)\})] \leq \mathbb{E}[\phi(\{Y(t), t \in [0, \infty)\})] \quad (2)$$

holds for all non-decreasing functional $\phi : \mathbb{V} \to \mathbb{R}$, provided that the expectations in (2) exist.

B. Queueing System Model

Consider the status update system that is illustrated in Fig. 1, where N flows of update packets are sent through a queue with M servers and an infinite buffer. Let s_n and d_n denote the source and destination nodes of flow n, respectively. Different flows can have different source and/or destination nodes. Each packet can be assigned to any server, and a server can only process one packet at a time. The service times of the update packets are *i.i.d.* across the servers and time.

The system starts to operate at time $t = 0$. The i-th update packet of flow n is generated at the source node s_n at time $S_{n,i}$, arrives at the queue at time $A_{n,i}$, and is delivered to the destination d_n at time $D_{n,i}$ such that $0 \leq S_{n,1} \leq S_{n,2} \leq \ldots$ and $S_{n,i} \leq A_{n,i} \leq D_{n,i}$. We consider the following class of *synchronized* packet generation and arrival processes:

Definition 4. *Synchronized Sampling and Arrivals:* The packet generation and arrival times are said to be *synchronized* across the N flows, if there exist two sequences $\{S_1, S_2, \ldots\}$ and $\{A_1, A_2, \ldots\}$ such that for all $i = 1, 2, \ldots$, and $n = 1, \ldots, N$

$$S_{n,i} = S_i, \ A_{n,i} = A_i. \quad (3)$$

Note that in this paper, the sequences $\{S_1, S_2, \ldots\}$ and $\{A_1, A_2, \ldots\}$ are *arbitrary*. Hence, *out-of-order arrivals*, e.g., $S_i < S_{i+1}$ but $A_i > A_{i+1}$, are allowed. In addition, when there is a single flow ($N = 1$), synchronized sampling and arrivals reduce to arbitrary packet generation and arrival processes that were considered in [5]–[8].

Let π represent a scheduling policy that determines the packet being sent by the servers over time. Let Π denote the set of *causal* policies in which the scheduling decisions are made based on the history and current states of the system. A policy is said to be *preemptive*, if each server can switch to send another packet at any time; the preempted packet will be stored back to the queue, waiting to be sent at a later time. A policy is said to be *non-preemptive*, if each server must complete sending the current packet before starting to serve another packet. A policy is said to be *work-conserving*, if all servers are kept busy whenever the queue is non-empty. We use Π_{np} to denote the set of non-preemptive causal policies such that $\Pi_{np} \subset \Pi$. Let

$$\mathcal{I} = \{S_i, A_i, \ i = 1, 2, \ldots\} \quad (4)$$

denote the packet generation and arrival times of the flows. We assume that the packet generation/arrival times $\mathcal{I}$ and the packet service times are determined by two *mutually independent* external processes, both of which do not change according to the adopted scheduling policy.

C. Age Metrics

At any time $t \geq 0$, the freshest packet delivered to the destination node d_n is generated at time

$$U_n(t) = \max\{S_{n,i} : D_{n,i} \leq t, i = 1, 2, \ldots\}. \quad (5)$$

The *age of information*, or simply the *age*, of flow n is defined as [1], [2]

$$\Delta_n(t) = t - U_n(t), \quad (6)$$

which is the time difference between the current time t and the generation time of the freshest packet currently available at destination d_n. Let $\mathbf{\Delta}(t) = (\Delta_1(t), \ldots, \Delta_N(t))$ denote the age vector of the N flows at time t.

We introduce an *age penalty function* $p(\mathbf{\Delta}) = p \circ \mathbf{\Delta}$ to represent the level of dissatisfaction for having aged information at the N destinations, where $p : \mathbb{R}^N \to \mathbb{R}$ can be any *non-decreasing* function of the N-dimensional age vector $\mathbf{\Delta}$. Some examples of the age penalty function are:

1. The *average age* of the N flows is

$$p_{\text{avg}}(\mathbf{\Delta}) = \frac{1}{N}\sum_{n=1}^{N} \Delta_n. \tag{7}$$

2. The *maximum age* of the N flows is

$$p_{\max}(\mathbf{\Delta}) = \max_{n=1,\ldots,N} \Delta_n. \tag{8}$$

3. The *mean square age* of the N flows is

$$p_{\text{ms}}(\mathbf{\Delta}) = \frac{1}{N}\sum_{n=1}^{N} (\Delta_n)^2. \tag{9}$$

4. The *l-norm of the age vector* of the N flows is

$$p_{l\text{-norm}}(\mathbf{\Delta}) = \left[\sum_{n=1}^{N} (\Delta_n)^l\right]^{\frac{1}{l}}, \ l \geq 1. \tag{10}$$

5. The *sum age penalty function* of the N flows is

$$p_{\text{sum-penalty}}(\mathbf{\Delta}) = \sum_{n=1}^{N} g(\Delta_n), \tag{11}$$

where $g : [0, \infty) \to \mathbb{R}$ is the age penalty function for each flow, which can be any *non-decreasing* function of the age Δ of the flow [3], [4]. For example, a stair-shape function $g_1(\Delta) = \lfloor a\Delta \rfloor$ with $a \geq 0$ can be used to characterize the dissatisfaction of data staleness when the information of interests is checked periodically, and an exponential function $g_2(\Delta) = e^{a\Delta}$ is appropriate for online learning and control applications where the desire for information refreshing grows quickly with respect to the age [4].

In this paper, we consider a class of *symmetric* and *non-decreasing* age penalty functions, i.e.,

$$\mathcal{P}_{\text{sym}} = \{p : [0, \infty)^N \to \mathbb{R} \text{ is symmetric and non-decreasing}\}.$$

This is a fairly large class of age penalty functions, where the function p can be discontinuous, non-convex, or non-separable. It is easy to see

$$\{p_{\text{avg}}, p_{\max}, p_{\text{ms}}, p_{l\text{-norm}}, p_{\text{sum-penalty}}\} \subset \mathcal{P}_{\text{sym}}.$$

Note that the age vector $\mathbf{\Delta}$ is a function of time t and policy π, and the age penalty function p may change over time. We use $\{p_t \circ \mathbf{\Delta}_\pi(t), t \in [0, \infty)\}$ to represent the stochastic process generated by the *time-dependent* age penalty function p_t in policy π. We assume that the initial age $\mathbf{\Delta}_\pi(0^-)$ at time $t = 0^-$ remains the same for all $\pi \in \Pi$.

IV. Multi-flow Update Scheduling

In this section, we investigate update scheduling of multiple information flows. We first consider a system setting with a single server and exponential service times, where an age optimality result is established. Next, we study a more general system setting with multiple servers and NBU service times. In this case, age optimality is inherently difficult to achieve and we present a near age-optimal result.

A. Multiple Flows, Single Server, Exponential Service Times

To address the multi-flow online scheduling problem, we consider a flow selection discipline called *Maximum Age First (MAF)* [12], [13], [17], in which *the flow with the maximum age is served the first, with ties broken arbitrarily*. A scheduling policy is defined by combining the MAF and LGFS disciplines as follows:

Definition 5. *Maximum Age First, Last Generated First Served (MAF-LGFS) policy:* In this policy, the last generated packet from the flow with the maximum age is served the first among all packets of all flows, with ties broken arbitrarily.

The age optimality of the preemptive MAF-LGFS policy is established in the following theorem.

Theorem 1. If (i) there is a single server ($M = 1$), (ii) the packet generation and arrival times are *synchronized* across the N flows, and (iii) the packet service times are exponentially distributed and *i.i.d.* across time, then it holds that for all $\mathcal{I}$, all $p_t \in \mathcal{P}_{\text{sym}}$, and all $\pi \in \Pi$

$$\begin{aligned}&[\{p_t \circ \mathbf{\Delta}_{\text{prmp, MAF-LGFS}}(t), t \geq 0\}|\mathcal{I}]\\ &\leq_{\text{st}}[\{p_t \circ \mathbf{\Delta}_\pi(t), t \geq 0\}|\mathcal{I}],\end{aligned} \tag{12}$$

or equivalently, for all $\mathcal{I}$, all $p_t \in \mathcal{P}_{\text{sym}}$, and all non-decreasing functional $\phi : \mathbb{V} \mapsto \mathbb{R}$

$$\begin{aligned}&\mathbb{E}\left[\phi(\{p_t \circ \mathbf{\Delta}_{\text{prmp, MAF-LGFS}}(t), t \geq 0\})|\mathcal{I}\right]\\ &= \min_{\pi \in \Pi} \mathbb{E}\left[\phi(\{p_t \circ \mathbf{\Delta}_\pi(t), t \geq 0\})|\mathcal{I}\right],\end{aligned} \tag{13}$$

provided that the expectations in (13) exist, where $\mathbb{V}$ is the set of Lebesgue measurable functions defined in (1).

Proof idea. If the packet generation and arrival times are synchronized across the flows, the preemptive MAF-LGFS policy satisfies the following property: When a packet is delivered to its destination, the flow with the maximum age before the packet delivery will have the minimum age among the N flows once the packet is delivered.[1] This is one key idea used in the proof. See Appendix A for the details. □

Theorem 1 tells us that, for all age penalty functions in $\mathcal{P}_{\text{sym}}$, all number of flows N, and all synchronized packet generation and arrival times $\mathcal{I}$, the preemptive MAF-LGFS policy minimizes the stochastic process $\{p_t \circ \mathbf{\Delta}_\pi(t), t \geq 0\}$ among all causal policies in a stochastic ordering sense. We note that a weaker version of Theorem 1 is to consider the mixture over the realizations of $\mathcal{I}$, where (12) becomes

$$[\{p_t \circ \mathbf{\Delta}_{\text{prmp, MAF-LGFS}}(t), t \geq 0\}] \leq_{\text{st}} [\{p_t \circ \mathbf{\Delta}_\pi(t), t \geq 0\}],$$

and similarly, the condition $\mathcal{I}$ in (13) can be removed.

B. Multiple Flows, Multiple Servers, NBU Service Times

Next, we consider a more general system setting with multiple servers and a class of New-Better-than-Used (NBU) service time distributions that include exponential distribution as a special case.

Definition 6. *New-Better-than-Used Distributions:* Consider a non-negative random variable X with complementary cumulative distribution function (CCDF) $\bar{F}(x) = \Pr[X > x]$. Then, X is said to be *New-Better-than-Used (NBU)* if for all $t, \tau \geq 0$

$$\bar{F}(\tau + t) \leq \bar{F}(\tau)\bar{F}(t). \tag{14}$$

[1] Note that this property does not hold when packet generations and arrivals are asynchronized.

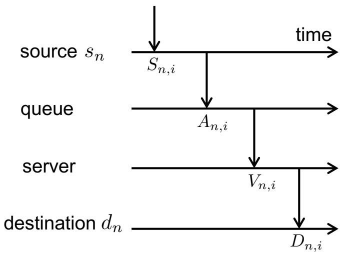

Fig. 2. An illustration of $S_{n,i}$, $A_{n,i}$, $V_{n,i}$, and $D_{n,i}$.

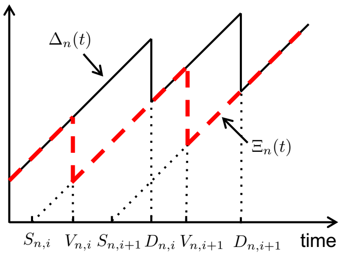

Fig. 3. The Age of Served Information $\Xi_n(t)$ as a lower bound of age $\Delta_n(t)$.

Examples of NBU distributions include constant service time, exponential distribution, shifted exponential distribution, geometrical distribution, Erlang distribution, negative binomial distribution, etc.

In scheduling literature, optimal online scheduling has been successfully established for delay minimization in single-server queueing systems, e.g., [18], [19], but can become inherently difficult in the multi-server cases. In particular, minimizing the average delay in deterministic scheduling problems with more than one servers is NP-hard [20]. Similarly, delay-optimal stochastic scheduling in multi-class, multi-server queueing systems is deemed to be notoriously difficult [21]–[23]. The key challenge in multi-class, multi-server scheduling is that *one cannot combine the resources of all the servers to jointly process the most important packet*. Due to the same reason, age-optimal online scheduling is quite challenging in multi-flow, multi-server systems. In the sequel, we consider a slightly relaxed goal to seek for *near* age-optimal online scheduling of multiple information flows, where our proposed scheduling policy is shown to be within a small additive gap from the optimum age performance.

Notice that the age $\Delta_n(t)$ in (6) is determined by the packets that have been delivered to the destination d_n by time t. To establish near age optimality, we consider an alternative age metric call the *Age of Served Information*, which is determined by the packets that have started service by time t: Let $V_{n,i}$ denote the time that the i-th packet of flow n is assigned to a server, i.e., the service starting time of the i-th packet of flow n, which is shown in Fig. 2. By definition, one can get $S_{n,i} \le A_{n,i} \le V_{n,i} \le D_{n,i}$. The *Age of Served Information* of flow n is defined as

$$\Xi_n(t) = t - \max\{S_{n,i} : V_{n,i} \le t, i = 1, 2, \ldots\}, \qquad (15)$$

which is the time difference between the current time t and the generation time of the freshest packet that has started service by time t. As shown in Fig. 3, $\Xi_n(t) \le \Delta_n(t)$. Let $\mathbf{\Xi}(t) = (\Xi_1(t), \ldots, \Xi_N(t))$ denote the Age of Served Information vector at time t.

We propose a new scheduling discipline called *Maximum Age of Served Information First (MASIF)*, in which *the flow with the maximum Age of Served Information is served the first, with ties broken arbitrarily*. Using this discipline, we define the following scheduling policy:

Definition 7. *Maximum Age of Served Information first, Last Generated First Served (MASIF-LGFS) policy:* In this policy, the last generated packet from the flow with the maximum *Age of Served Information* is served the first among all packets of all flows, with ties broken arbitrarily.

In some previous studies, e.g., [12], [13], [24], it was proposed to discard old packets and only store and send the freshest one. While this technique can reduce the age, in many applications such as social updates, news seeds, and stock trading, some old packets with earlier generation times are still quite useful and are needed to be sent to the destinations. Next, we will show that the non-preemptive MASIF-LGFS policy, which does not discard old packets, is near age-optimal. Hence, the additional age reduction provided by discarding old packets in the non-preemptive MASIF-LGFS policy is not large. In order to establish this result, we first show that the age of served information of the non-preemptive MASIF-LGFS policy provides a uniform age lower bound for all non-preemptive and causal policies.

Theorem 2. If (i) the packet generation and arrival times are *synchronized* across the N flows and (ii) the packet service times are NBU and *i.i.d.* across the servers and time, then it holds that for all $\mathcal{I}$, all $p_t \in \mathcal{P}_{\text{sym}}$, and all $\pi \in \Pi_{np}$

$$\begin{aligned}&[\{p_t \circ \mathbf{\Xi}_{\text{non-prmp, MASIF-LGFS}}(t), t \ge 0\}|\mathcal{I}]\\ \le_{\text{st}}&[\{p_t \circ \mathbf{\Delta}_\pi(t), t \ge 0\}|\mathcal{I}], \qquad (16)\end{aligned}$$

or equivalently, for all $\mathcal{I}$, all $p_t \in \mathcal{P}_{\text{sym}}$, and all non-decreasing functional $\phi : \mathbb{V} \mapsto \mathbb{R}$

$$\begin{aligned}&\mathbb{E}\left[\phi(\{p_t \circ \mathbf{\Xi}_{\text{non-prmp, MASIF-LGFS}}(t), t \ge 0\})|\mathcal{I}\right]\\ \le&\min_{\pi \in \Pi_{np}} \mathbb{E}\left[\phi(\{p_t \circ \mathbf{\Delta}_\pi(t), t \ge 0\})|\mathcal{I}\right]\\ \le&\mathbb{E}\left[\phi(\{p_t \circ \mathbf{\Delta}_{\text{non-prmp, MASIF-LGFS}}(t), t \ge 0\})|\mathcal{I}\right], \qquad (17)\end{aligned}$$

provided that the expectations in (17) exist.

Proof idea. Under synchronized packet generations and arrivals, the non-preemptive MASIF-LGFS policy satisfies: When a packet starts service, the flow with the maximum Age of Served Information before the service starts will have the minimum Age of Served Information among the N flows once the service starts. Theorem 2 is proven by using this idea and the sample-path method developed in [25], [26]. We note that the sample-path method in [25], [26] is the key for addressing the challenge in multi-flow, multi-server scheduling. See our technical report [27] and [25], [26] for the details. □

Hence, the non-preemptive MASIF-LGFS policy is near age-optimal in the sense of (17). In particular, for the average age of the N flows in (7) (i.e., $p_t = p_{\text{avg}}$), we can obtain

Theorem 3. Under the conditions of Theorem 2, it holds that for all $\mathcal{I}$

$$\min_{\pi \in \Pi_{np}} [\bar{\Delta}_\pi|\mathcal{I}] \le [\bar{\Delta}_{\text{non-prmp, MASIF-LGFS}}|\mathcal{I}] \le \min_{\pi \in \Pi_{np}} [\bar{\Delta}_\pi|\mathcal{I}] + \mathbb{E}[X],$$

where $[\bar{\Delta}_\pi|\mathcal{I}] = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}[\int_0^T p_{\text{avg}} \circ \mathbf{\Delta}_\pi(t)dt|\mathcal{I}]$ is the expected time-average of the average age of the N flows, and $\mathbb{E}[X]$ is the mean service time of one packet.

The proof of Theorem 3 is similar to that of Theorem 4 in [6] and hence is omitted here. By Theorem 3, the average

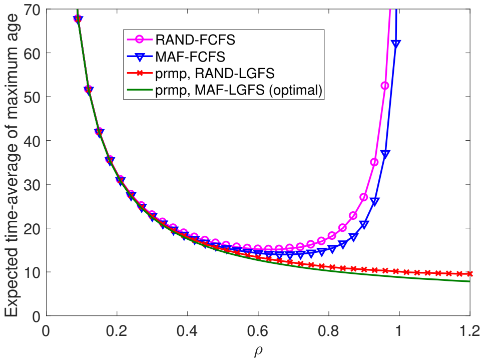

Fig. 4. Expected time-average of the maximum age of 3 flows in a system with a single server and *i.i.d.* exponential service times.

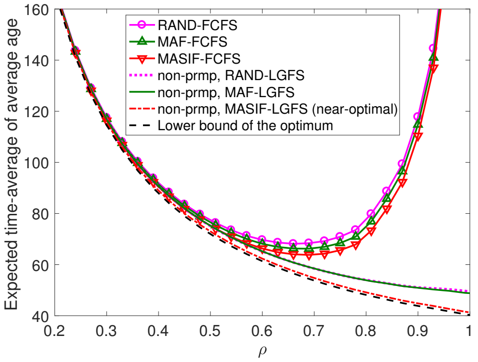

Fig. 5. Expected time-average of the average age of 50 flows in a system with 3 servers and *i.i.d.* NBU service times.

age of the non-preemptive MASIF-LGFS policy is within an additive gap from the optimum, and the gap $\mathbb{E}[X]$ is invariant of the packet arrival and generation times $\mathcal{I}$, the number of flows N, and the number of servers M.

V. Numerical Results

In this section, we evaluate the age performance of several multi-flow online scheduling policies. These scheduling policies are defined by combining the flow selection disciplines $\{$MAF, MASIF, RAND$\}$ and the packet selection disciplines $\{$FCFS, LGFS$\}$, where RAND represents randomly choosing a flow among the flows with un-served packets. The packet generation times S_i follow a Poisson process with rate λ, and the time difference $(A_i - S_i)$ between packet generation and arrival is equal to either 0 or $4/\lambda$ with equal probability. The mean service time of each server is set as $\mathbb{E}[X] = 1/\mu = 1$. Hence, the traffic intensity is $\rho = \lambda N/M$, where N is the number of flows and M is the number of servers.

Figure 4 illustrates the expected time-average of the maximum age $p_{\max}(\boldsymbol{\Delta}(t))$ of 3 flows in a system with a single server and *i.i.d.* exponential service times. One can see that the preemptive MAF-LGFS policy has the best age performance and its age is quite small even for $\rho > 1$, in which case the queue is actually unstable. On the other hand, both the RAND and FCFS disciplines have much higher age. Note that there is no need for preemptions under the FCFS discipline. Figure 5 plots the expected time-average of the average age $p_{\text{avg}}(\boldsymbol{\Delta}(t))$ of 50 flows in a system with 3 servers and *i.i.d.* NBU service times. In particular, the service time X follows the following shifted exponential distribution:

$$\Pr[X > x] = \begin{cases} 1, & \text{if } x < \frac{1}{3}; \\ \exp[-\frac{3}{2}(x - \frac{1}{3})], & \text{if } x \geq \frac{1}{3}. \end{cases} \quad (18)$$

One can observe that the non-preemptive MASIF-LGFS policy is better than the other policies, and is quite close to the age lower bound where the gap from the lower bound is no more than the mean service time $\mathbb{E}[X] = 1$. One interesting observation is that the non-preemptive MASIF-LGFS policy is better than the non-preemptive MAF-LGFS policy for NBU service times. The reason behind this is as follows: When multiple servers are idle in the non-preemptive MAF-LGFS policy, these servers are assigned to process multiple packets from the flow with the maximum age (say flow n). This reduces the age of flow n, but at a cost of postponing the service of the flows with the second and third maximum ages. On the other hand, in the non-preemptive MASIF-LGFS policy, once a packet from the flow with the maximum age of served information (say flow m) is assigned to a server, the age of served information of flow m drops greatly and the next server will be assigned to process the flow with second maximum age of served information. As shown in [25], [26], using multiple parallel servers to process different flows is beneficial for NBU service times. The behavior of non-preemptive MASIF-LGFS policy is similar to the maximum matching scheduling algorithms, e.g., [28], [29] for time-slotted systems, where multiple servers are assigned to process different flows in each time-slot. One difference is that the non-preemptive MASIF-LGFS policy can even operate in continuous-time systems, but the maximum matching scheduling algorithms cannot.

VI. Conclusion and Future Work

We have developed online scheduling policies and shown they are (near) optimal for minimizing the age of information in multi-flow, multi-server systems. Similar with [6], the results in this paper can be generalized to consider packet replications over multiple servers. In addition, similar to the results in [25], [26], Theorem 2 and Theorem 3 can be generalized to the case that the servers have different NBU service time distributions. Other future research directions include asynchronized packet arrivals, packet transmissions with errors, and multi-flow updates in multi-hop networks.

References

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

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

[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," in *IEEE INFOCOM*, 2016.

[4] ——, "Update or wait: How to keep your data fresh," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7492–7508, Nov. 2017.

[5] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Optimizing data freshness, throughput, and delay in multi-server information-update systems," in *IEEE ISIT*, 2016.

[6] ——, "Minimizing the age of information through queues," submitted to *IEEE Trans. Inf. Theory*, 2017, http://arxiv.org/abs/1709.04956.

[7] ——, "Age-optimal information updates in multihop networks," in *IEEE ISIT*, 2017.

[8] ——, "The age of information in multihop networks," submitted to *IEEE Trans. Inf. Theory*, 2017, https://arxiv.org/abs/1712.10061.

[9] R. D. Yates and S. K. Kaul, "Real-time status updating: Multiple sources," in *IEEE ISIT*, July 2012, pp. 2666–2670.

[10] ——, "The age of information: Real-time status updating by multiple sources," submitted to *IEEE Trans. Inf. Theory*, 2016, http://arxiv.org/abs/1608.08622.

[11] L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *Proc. IEEE ISIT*, June 2015, pp. 1681–1685.

[12] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, “Minimizing age of information in broadcast wireless networks,” in *Allerton Conf. on Communication, Control, and Computing*, September 2016.

[13] Y.-P. Hsu, E. Modiano, and L. Dua, “Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals,” submitted to *IEEE Trans. Wireless Commun.*, 2017, https://arxiv.org/abs/1712.07419.

[14] Q. He, D. Yuan, and A. Ephremides, “Optimal link scheduling for age minimization in wireless systems,” *IEEE Trans. Inf. Theory*, in press, 2018.

[15] R. D. Yates and S. K. Kaul, “Status updates over unreliable multiaccess channels,” in *IEEE ISIT*, 2017, pp. 331–335.

[16] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*. Springer, 2007.

[17] B. Li, A. Eryilmaz, and R. Srikant, “On the universality of age-based scheduling in wireless networks,” in *IEEE INFOCOM*, April 2015, pp. 1302–1310.

[18] L. Schrage, “A proof of the optimality of the shortest remaining processing time discipline,” *Operations Research*, vol. 16, pp. 687–690, 1968.

[19] J. R. Jackson, “Scheduling a production line to minimize maximum tardiness,” management Science Research Report, University of California, Los Angeles, CA, 1955.

[20] S. Leonardi and D. Raz, “Approximating total flow time on parallel machines,” in *ACM STOC*, 1997.

[21] G. Weiss, “Turnpike optimality of Smith’s rule in parallel machines stochastic scheduling,” *Math. Oper. Res.*, vol. 17, no. 2, pp. 255–270, May 1992.

[22] ——, “On almost optimal priority rules for preemptive scheduling of stochastic jobs on parallel machines,” *Advances in Applied Probability*, vol. 27, no. 3, pp. 821–839, 1995.

[23] M. Dacre, K. Glazebrook, and J. Nio-Mora, “The achievable region approach to the optimal control of stochastic systems,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 747–791, 1999.

[24] M. Costa, M. Codreanu, and A. Ephremides, “On the age of information in status update systems with packet management,” *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1897–1910, April 2016.

[25] Y. Sun, C. E. Koksal, and N. B. Shroff, “On delay-optimal scheduling in queueing systems with replications,” 2016, https://arxiv.org/abs/1603.07322.

[26] ——, “Near delay-optimal scheduling of batch jobs in multi-server systems,” 2017, http://www.auburn.edu/%7eyzs0078/parallel-servers.pdf.

[27] Y. Sun, E. Uysal-Biyikoglu, and S. Kompella, “Age-optimal updates of multiple information flows,” technical report, 2018, https://arxiv.org/abs/1801.02394.

[28] C. Joo, X. Lin, and N. B. Shroff, “Understanding the capacity region of the greedy maximal scheduling algorithm in multihop wireless networks,” *IEEE/ACM Trans. Netw.*, vol. 17, no. 4, pp. 1132–1145, Aug. 2009.

[29] B. Ji, G. R. Gupta, X. Lin, and N. B. Shroff, “Low-complexity scheduling policies for achieving throughput and asymptotic delay optimality in multichannel wireless networks,” *IEEE/ACM Transactions on Networking*, vol. 22, no. 6, pp. 1911–1924, Dec 2014.

Appendix A
Proof of Theorem 1

We first establish two lemmas that are useful to prove Theorem 1. Let the age vector $\mathbf{\Delta}_{\pi}(t)$ denote the *system state* of policy π at time t and $\{\mathbf{\Delta}_{\pi}(t), t \geq 0\}$ denote the *state process* of policy π. For notational simplicity, let policy P represent the preemptive MAF-LGFS policy. Using the memoryless property of exponential distribution, we can obtain the following coupling lemma:

Lemma 1. *(Coupling Lemma)* For any given $\mathcal{I}$, consider policy P and any *work-conserving* policy $\pi \in \Pi$. If (i) there is a single server ($M = 1$) and (ii) the packet service times are exponentially distributed and *i.i.d.* across time, then there exist policy P_1 and policy π_1 in the same probability space which satisfy the same scheduling disciplines with policy P and policy π, respectively, such that

1. The state process $\{\mathbf{\Delta}_{P_1}(t), t \geq 0\}$ of policy P_1 has the same distribution with the state process $\{\mathbf{\Delta}_{P}(t), t \geq 0\}$ of policy P,
2. The state process $\{\mathbf{\Delta}_{\pi_1}(t), t \geq 0\}$ of policy π_1 has the same distribution with the state process $\{\mathbf{\Delta}_{\pi}(t), t \geq 0\}$ of policy π,
3. If a packet is delivered at time t in policy P_1 as $\mathbf{\Delta}_{P_1}(t)$ evolves, then almost surely, a packet is delivered at time t in policy π_1 as $\mathbf{\Delta}_{\pi_1}(t)$ evolves; and vice versa.

Proof. See our technical report [27]. □

We will compare policy P_1 and policy π_1 on a sample path by using the following lemma:

Lemma 2. *(Inductive Comparison)* Under the conditions of Lemma 1, suppose that a packet is delivered in policy P_1 and a packet is delivered in policy π_1 at the same time t. The system state of policy P_1 is $\mathbf{\Delta}_{P_1}$ before the packet delivery, which becomes $\mathbf{\Delta}'_{P_1}$ after the packet delivery. The system state of policy π_1 is $\mathbf{\Delta}_{\pi_1}$ before the packet delivery, which becomes $\mathbf{\Delta}'_{\pi_1}$ after the packet delivery. If the packet generation and arrival times are *synchronized* across the N flows and

$$\Delta_{[i],P_1} \leq \Delta_{[i],\pi_1},\ i = 1, \ldots, N, \tag{19}$$

then

$$\Delta'_{[i],P_1} \leq \Delta'_{[i],\pi_1},\ i = 1, \ldots, N. \tag{20}$$

Proof. See our technical report [27]. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. Consider any work-conserving policy $\pi \in \Pi$. By Lemma 1, there exist policy P_1 and policy π_1 satisfying the same scheduling disciplines with policy P and policy π, respectively, and the packet delivery times in policy P_1 and policy π_1 are synchronized almost surely.

For any given sample path of policy P_1 and policy π_1, $\mathbf{\Delta}_{P_1}(0^-) = \mathbf{\Delta}_{\pi_1}(0^-)$ at time $t = 0^-$. We consider two cases:

Case 1: When there is no packet delivery, the age of each flow grows linearly with a slope 1.

Case 2: When a packet is delivered, the evolution of the age is governed by Lemma 2.

By induction over time, we obtain

$$\Delta_{[i],P_1}(t) \leq \Delta_{[i],\pi_1}(t),\ i = 1, \ldots, N,\ t \geq 0. \tag{21}$$

For any symmetric and non-decreasing function p_t, it holds from (21) that for all sample paths and all $t \geq 0$

$$\begin{aligned} & p_t \circ \mathbf{\Delta}_{P_1}(t) \\ =& p_t(\Delta_{1,P_1}(t), \ldots, \Delta_{N,P_1}(t)) \\ =& p_t(\Delta_{[1],P_1}(t), \ldots, \Delta_{[N],P_1}(t)) \\ \leq& p_t(\Delta_{[1],\pi_1}(t), \ldots, \Delta_{[N],\pi_1}(t)) \\ =& p_t(\Delta_{1,\pi_1}(t), \ldots, \Delta_{N,\pi_1}(t)) \\ =& p_t \circ \mathbf{\Delta}_{\pi_1}(t). \end{aligned} \tag{22}$$

By Lemma 1, the state process $\{\mathbf{\Delta}_{P_1}(t), t \geq 0\}$ of policy P_1 has the same distribution with the state process $\{\mathbf{\Delta}_{P}(t), t \geq 0\}$ of policy P; the state process $\{\mathbf{\Delta}_{\pi_1}(t), t \geq 0\}$ of policy π_1 has the same distribution with the state process $\{\mathbf{\Delta}_{\pi}(t), t \geq 0\}$ of policy π. By (22) and Theorem 6.B.30 in [16], (12) holds for all work-conserving policy $\pi \in \Pi$.

For non-work-conserving policies π, because the service times are exponentially distributed and *i.i.d.* across servers and time, server idling only postpones the delivery times of the packets. One can construct a coupling to show that for any non-work-conserving policy π, there exists a work-conserving policy π' whose age process is smaller than that of policy π in stochastic ordering; the details are omitted. As a result, (12) holds for all policies $\pi \in \Pi$.

Finally, the equivalence between (12) and (13) follows from (2). This completes the proof. □