Wireless Scheduling for Information Freshness and Synchrony: Drift-based Design and Heavy-Traffic Analysis

Changhee Joo
Electrical and Computer Engineering
UNIST
Ulsan, 44919, Korea
cjoo@unist.ac.kr

Atilla Eryilmaz
Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43210, USA
eryilmaz.2@osu.edu

***Abstract*—We consider the problem of scheduling in wireless networks with the aim of maintaining up-to-date and synchronized (also called, *aligned*) information at the receiver across multiple flows. This is in contrast to the more conventional approach of scheduling for optimizing long-term performance metrics such as throughput, fairness, or average delay. Maintaining the age of information at a low and roughly equal level is particularly important for distributed cyber-physical systems, in which the effectiveness of the control decisions depends critically on the freshness and synchrony of information from multiple sources/sensors. In this work, we first expose the weakness of several popular MaxWeight scheduling solutions that utilize queue-length, delay, and age information as their weights. Then, we develop a novel age-based scheduler that combines age with the interarrival times of incoming packets in its decisions, which yields significant gains in the information freshness at the receiver. We characterize the performance of our strategy through a heavy-traffic analysis that establishes upper and lower bounds on the freshness of system information.**

I. Introduction

Wireless networks are expected to form the communication backbone of many future cyber-physical systems that are expected to support diverse applications such as autonomous driving in vehicular networks, monitoring and response in sensor networks, efficient supply and demand management in smart power grids, etc. As such, wireless networks are no longer merely a medium of high-rate information transfer that are detached from the content of the information, but an integral part of a distributed controller-actuator system whose performance is highly dependent on the timeliness and accuracy of the information that guides the system operation.

Over the last few decades, wireless resource allocation research has been increasingly more effective in maximizing long-term performance metrics such as throughput, utility, reliability, and average delay (see [1]–[6] and references therein). These advances have benefited from an ever-expanding framework of *adaptive* controller design that utilize measures such as actual/virtual queue-length (e.g., [7]–[13]), Head-of-Line (HoL) delay (e.g., [14]–[18]), drop-rates (e.g., [19], [20]), time-since-last-service (e.g., [21], [22]) information in order to guide a variety of decisions including rate control, scheduling, and routing.

Separate from these developments, relatively recently there has been an interest in maintaining *fresh* information of a flow at the receiving end of a communication link (e.g., [23]–[27]). This is important in applications, where the freshness of the system state information is critical to the control decisions. Most of these prior works (e.g., [23]–[26]) have focused on the analysis and/or control of the status updates from a single source or multiple sources to a single server, i.e., maintaining up-to-date 'status of the source(s)' at the receiver.

In this paper, we consider a different concept of freshness that is measured by the 'age of received packet' from each source. Such a measure is motivated by applications where it is important to maintain *equally delayed* information from multiple sources at the receiver, such as network monitoring and distributed sensing. In these applications, it is important that the information flow from different sources are roughly synchronized for accurate tracking (in monitoring applications) and stable control (in distributed sensing and control applications). Further, we consider the problem of scheduling for fresh information in a general network of wirelessly inter-connected servers that receive randomly arriving stream of updates. This setting calls for a different set of models as well as analysis and design tools than those employed in the aforementioned works. The more recent work [27] considers the problem of scheduling for fresh information in wireless networks, and presents a set of interesting structural results concerning the tractability and intractability of the optimal scheduling solution. It also provides a so-called steepest-age-descent algorithm that is numerically investigated. In our work, we take a different approach based on the drift-minimization methodology, and conduct a heavy-traffic analysis of its performance in terms of the freshness metric. We believe that these complementary works collectively help expand our understanding and management of networks for the new metric of information freshness.

With this vision, we first provide a measure of information freshness for multi-source wireless networks based on a virtual queueing model. Then, we present a comparative investigation of three well-known scheduling strategies – namely, two MaxWeight Schedulers that use queue-lengths and HoL delays as their weights, and a round-robin scheduler – to reveal that each of these three choices can result in deficient scheduling choices for the new freshness metric.

This work is supported in part by IITP grant funded by the Korea government (MSIP) No. B0126-16-1064; Research on Near-Zero Latency Network for 5G Immersive Service). the NSF grants: CCSS-EARS-1444026, CNS-NeTS-1514127, CMMI-SMOR-1562065 and CNS-WiFiUS-1456806, CNS-ICN-WEN-1719371; the DTRA grant HDTRA1-15-1-0003. The work of A. Eryilmaz is also supported by the QNRF Grant NPRP 7-923-2-344.

Based on these observations, we develop a new age-based scheduler that combines age information with interarrival times in order to determine the weights assigned to different flows. To characterize the performance of our proposed scheduler, we also perform its heavy-traffic analysis that yields lower and upper bounds on the heavy-traffic performance of our proposed policy. Heavy-traffic analysis has been an effective methodology for analyzing the performance of scheduling policies (e.g., see [28] and references therein). While the results are obtained under heavy-traffic conditions, the scheduler possess desirable freshness characteristics even in lightly-loaded conditions, thereby making it a good choice for maintaining up-to-date information of flows at the receiving end.

The key message that we learn from this work is the value of interarrival times in maintaining fresh and equally-delayed information updates of continuous flows. This insight is expected to be useful in designing the communication backbone of future cyber-physical systems whose operation is critically dependent on freshness of information.

II. System Model

We consider a network graph $G = (N, L)$ with the set N of nodes and the set L of wireless links. Due to wireless interference, at each time t, a subset of links $\mathbf{S}(t) \in L$ can be scheduled at the same time. The subsets of links that satisfy the interference constraints are said to be a *feasible* schedule. Let $\mathcal{S}$ denote the set of all feasible schedules. Once a link is scheduled, it transmits one packet during the time slot. We assume that all the transmissions occur in a time slotted manner: the i-th packet at link l arrives at $t_{l,i} \in \mathbb{R}$ and is served at $t'_{l,i} \in \mathbb{N}$, where $\mathbb{R}$ denotes the set of real numbers and $\mathbb{N}$ denotes the set of non-negative integers.

At each link l, packets arrive following a stochastic process with mean rate λ_l. Let $\boldsymbol{\lambda}$ denote its vector. Also let $X_{l,i}$ denote the interarrival time between the i-th packet and the $(i+1)$-th packet at link l. We assume that the interarrival times are independent and bounded by X_{max}, i.e.,

$$X_{l,i} := t_{l,i+1} - t_{l,i} \leq X_{max}. \tag{1}$$

Let $A_l(t)$ denote the number of packet arrivals in $(t, t+1]$ for $t \in \mathbb{N}$, and let $S_l(t) \in \{0, 1\}$ denote the number of served packets in $(t, t+1]$, in particular, $S_l(t) = 1$ if $l \in \mathbf{S}(t)$ and $S_l(t) = 0$ if $l \notin \mathbf{S}(t)$. We slightly abuse our notation and use interchangeably the set $\mathbf{S}(t)$ of scheduled links and the vector $\{S_l(t)\}$ of served packets. Let $Q_l(t)$ denote the queue length at link l, which evolves as $Q_l(t+1) = (Q_l(t) - S_l(t))^+ + A_l(t)$, where $(\cdot)^+ := \max\{0, \cdot\}$. All the queues are served in a first-come-first-served manner. Let $N_l(t)$ denote the index of Head-of-Line (HoL) packet at the queue of link l at the beginning of time t, i.e.,

$$N_l(t) := \min\{i \mid t'_{l,i} \geq t\}, \tag{2}$$

which is well-defined when $Q_l(t) > 0$.

We define the *age* of link l as the difference between the current time and the time when the HoL packet of link l is generated. The age is set to 0 if the queue is empty. As such, age is a measure of how *outdated* the data at the receiving end of the link is compared to the data at its transmitting end. Assuming that only the links with non-zero queue can be scheduled, the age $U_l(t)$ of link l can be considered as a virtual queue that evolves as, for $t \in \mathbb{N}$,

$$U_l(t+1) = \begin{cases} U_l(t) + \mathbb{1}_{\{Q_l(t)>0\}}, & \text{if } l \notin \mathbf{S}(t), \\ \big(U_l(t) + 1 - X_{l,N_l(t)}\big)^+, & \text{if } l \in \mathbf{S}(t), \end{cases} \tag{3}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indication function. The first equation implies that the age increases by 1 when the packet is not served, and the second equation implies that the age decreases by the amount of interarrival time when the packet is served. We note that our definition of the age is slightly different from [23]. Specifically, the age equals 0 when the queue is empty under our definition, and accounts for the oldness of the information waiting at the HoL of the link. Also, we assume $Q_l(0) = 0$ and $U_l(0) = 0$.

We say that the system is *stable* if the time-averaged mean ages of all the links remain finite. Let Λ denote the set of arrival rates such that for any $\boldsymbol{\lambda} \in \Lambda$ (strictly inside), there exists a scheduling policy that can stabilize the system. Note that from the Little's law, the stability region of age is equivalent to the stability region of the queue lengths, and any throughput-optimal schedulers that keep all the queue lengths finite (e.g., Queue-length based MaxWeight [7]) is an optimal solution that achieves Λ.

III. Motivation

In this section, we expose the deficiency of a *round-robin* scheduler as well as commonly used throughput-optimal MaxWeight schedulers that utilize *queue-lengths* and *delays* to make the scheduling decisions. In particular, we design flows with particular arrival patterns, and show that these popular schedulers are unable to keep system information freshness equally low. This will motivate us in the next section to develop and analyze a new *age-based* scheduler that is aimed at optimizing freshness of information.

Let us consider a simple network with three flows. Three flows A, B, C have deterministic packet arrivals of different patterns, and share a server that can serve one packet from one flow at a time. In this example, we assume that all packets arrive at the beginning of the time slot, and in each flow, packets are served in the first-come-first-serve manner. At time 1, all the three flows have a packet arrival. Flow A has additional arrival at time 3. The pattern repeats as shown in Fig. 1, where the k-th packet from flow Z is marked as Zk.

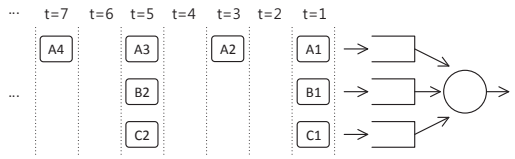


Fig. 1. Example of deterministic packet arrivals.

Suppose that there is no service until time slot 4 and we start transmitting the packets from time slot 5. First, we transmit the packets following the largest queue-length first policy. At time

5, we have the queue length vector of $\mathbf{Q} = \{Q_A, Q_B, Q_C\} = \{3, 2, 2\}$, and we serve A1. At time 6, we have $\mathbf{Q} = \{2, 2, 2\}$, and break the tie by transmitting the oldest packet first, i.e., B2 (or C2). At time 7, packet A4 arrives and we have $\mathbf{Q} = \{3, 1, 2\}$, and transmit A2. At time 8, we have $\mathbf{Q} = \{2, 1, 2\}$ and transmit C2. At time 9, we have $\mathbf{Q} = \{3, 2, 2\}$ and transmit a packet from flow A. It can be easily observed that the service repeats in the order of $\{A, B, A, C\}$.

Second, we consider another scheduling following the oldest packet first policy. At time 5, we have the flow age vector of $\mathbf{U} = \{U_A, U_B, U_C\} = \{4, 4, 4\}$. We break the tie in the order of {A,B,C}, and transmit A1. At time 6, we have $\mathbf{U} = \{3, 5, 5\}$ and transmit B1. At time 7, we have $\mathbf{U} = \{4, 2, 6\}$ and transmit C1. At time 8, we have $\mathbf{U} = \{5, 3, 3\}$ and transmit A2. At time 9, we have $\mathbf{U} = \{4, 4, 4\}$ and transmit a packet from flow A. It can be easily observed that the service repeats in the order of $\{A, B, C, A\}$.

Finally, we consider the scheduler that serves the packet with the largest age-weighted age drop. Let us consider the ages $(t - t_{l,N_l(t)}, t - t_{l,N_l(t)+1})$ of two packets at the head of queue. For example, at time 5, flow A has $(4, 2)$ for the age of two HoL packets, i.e., for (A1, A2), and flows B and C have $(4, 0)$. The scheduler chooses the packet that leads to the largest age-weighted age drop, i.e., the HoL packet of the flow with the largest $(t - t_{l,N_l(t)}) \cdot (t_{l,N_l(t)+1} - t_{l,N_l(t)})$ (break the tie in the order of {A,B,C}), and thus, we will schedule B1 at time 5. At time 6, the age of all the packets remained in the queues increases by one, and we have $\{(5, 3), (1, 0), (5, 1)\}$, where we set the age of not-yet-arrived packet to 0. The scheduler will transmit C1. At time 7, we have $\{(6, 4), (2, 0), (2, 0)\}$ and serve A1. At time 8, we have $\{(5, 3), (3, 0), (3, 0)$ and serve A2. At time 9, we have $\{(4, 2), (4, 0), (4, 0)\}$ and serve a packet from flow B. It can be easily observed that the service repeats in the order of $\{B, C, A, A\}$.

Under each scheduling policy, the packet delay of {A1,A2,B1,C1} can be calculated as in Table I. We first note that the total delay sums are equal for all the policies. In fact, it will be the same for all work-conserving schedulers. Then, we observe that they have different per-flow delays. Under the largest-queue-first policy, we have $\{4, 5, 7\}$ for flows A, B, C, respectively. Under the oldest-packet-first policy, we have $\{4.5, 5, 6\}$, and under the largest-age-weighted-drop-first policy, we have $\{5.5, 4, 5\}$.

TABLE I
DELAY OF EACH PACKET $(t'_{l,i} - t_{l,i})$ IN TIME SLOTS.

	largest-queue-first	oldest-packet-first	largest-age-weighted-drop-first
(A1,A2)	(4,4)	(4,5)	(6,5)
B1	5	5	4
C1	7	6	5

The result raises an interesting question about the fairness of packet delays, in particular when the flows have different arrival rates. Considering the oldest-packet-first policy and the largest-age-weighted-drop-first policy, they have similar per-flow delay performances, but they do have different preference, which will be clarified later in Section V. Prioritizing the packets (or information) of the same age with their flow's interarrival time can motivate the sources to decrease their transmission rate to achieve better delay performance. To this end, it is interesting to investigate how the ages related to the per-flow delay performance. Extending the largest-queue-first and the oldest-packet-first schemes, we introduce the well-known scheduling policies, and investigate their behaviors under symmetric and asymmetric traffic.

The solution that finds the schedule with the maximum queue-weighted sum, denoted by Q-MW, has been well-known to be throughput-optimal. At each time slot, it has the schedule $\mathbf{S}^Q(t)$ as

$$\mathbf{S}^Q(t) = \underset{\mathbf{S}\in\mathcal{S}}{\operatorname{argmax}} \sum_{l\in L} Q_l(t) \cdot S_l, \tag{4}$$

Another well-known throughput-optimal solution is the maximum HoL delay weighted sum, denoted by D-MW [17]. At each time slot, it has the schedule $\mathbf{S}^D(t)$ as

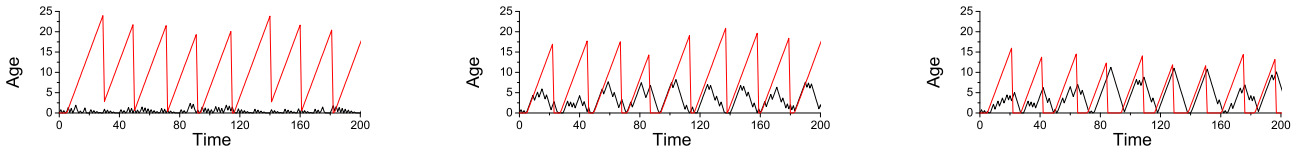
$$\mathbf{S}^D(t) = \underset{\mathbf{S}\in\mathcal{S}}{\operatorname{argmax}} \sum_{l\in L} U_l(t) \cdot S_l. \tag{5}$$

Also, the round-robin scheduler (RR) is a well-known alternative. Through simulations under simple scenarios, we demonstrate the age performance of these three scheduling choices.

We consider a symmetric scenario with two links. Each link has an on-off channel and turns on with probability 0.9, independently across times and links. Each link has a flow with the same mean packet arrival rate[1] 0.45, but their interarrival times are different. For one flow (regular flow), packets arrive with a fixed interarrival time, and for the other flow (bursty flow), packets arrive in a burst: 10 packets within 0.1 slot time. Note that all packets in a burst have similar generation times, and thus, the HoL delay of the link will keep increasing until all the packets in the burst are served out.

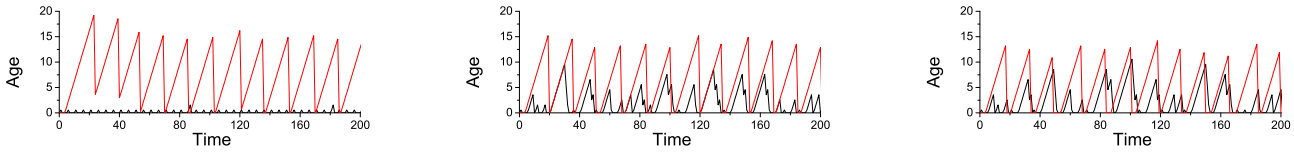
Fig. 2 shows the ages of the two flows (i.e. the HoL delay of the two links) under RR, Q-MW, and D-MW scheduling schemes, respectively. Under RR, the regular flow achieves good age performance while the bursty flow suffers from large ages. This is because the last packet of a burst has to wait for long time under RR. (Each age drop of the bursty flow indicates that the last packet of a burst is served out.) Under Q-MW, we can observe the age of the regular flow increases from when a packet burst of the bursty flow arrives. It is because the larger queue will be served first under Q-MW. Upon the arrival of a burst, the bursty flow will be served first, and then when the queue lengths of the two flows are the same, they will be served in turn. The priority given to the bursty flow reduces the age of the bursty flow less than that under RR. Under D-MW, the bursty flow has a priority if its burst arrive earlier than the HoL packet of the regular flow, which delays the packets of the regular flow and causes it to have as high ages as the bursty flow.

[1]The arrival rates are within the stability region, since the total arrival rate 0.9 is less than the channel opportunity rate $1 - 0.1^2 = 0.99$.



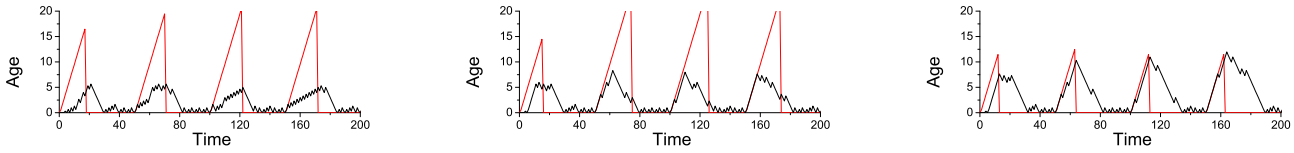
(a) RR (b) Q-MW (c) D-MW

Fig. 2. Ages of two flows with the same arrival rate $\lambda_{\text{regular}} = \lambda_{\text{bursty}}$. One flow has regular traffic (black) while the other has bursty traffic (red).



(a) RR (b) Q-MW (c) D-MW

Fig. 3. Ages of two flows with different arrival rates. $\lambda_{\text{regular}} = \lambda_{\text{bursty}}/3$.



(a) RR (b) Q-MW (c) D-MW

Fig. 4. Ages of two flows with different arrival rates. $\lambda_{\text{regular}} = 3\lambda_{\text{bursty}}$.

Similar results are observed when the bursty flow has a higher arrival rate than that of the regular traffic as shown in Fig. 3, where we set $\lambda_{\text{regular}} = \lambda_{\text{bursty}}/3 = 0.2$. However, when the bursty flow has a lower arrival rate, where we set $\lambda_{\text{regular}} = 3\lambda_{\text{bursty}} = 0.6$, we can observe that Q-MW suffers from large ages, as shown in Fig. 4.

TABLE II
TOTAL AVERAGE AGE

	RR	Q-MW	D-MW
$\lambda_{\text{regular}} = \lambda_{\text{bursty}}$	12.08	12.11	9.33
$\lambda_{\text{regular}} = \lambda_{\text{bursty}}/3$	7.64	8.27	7.20
$\lambda_{\text{regular}} = 3\lambda_{\text{bursty}}$	6.54	8.49	5.65

For each scenario, the total average age $\frac{1}{T}\sum_{\tau=1}^{T}\sum_l U_l(\tau)$ is as shown in Table II. It clarifies that Q-MW has the largest average age. An interesting result is that when the bursty flow has a lower arrival rate, the ages under Q-MW are larger than the ages under RR for *both the flows*: 5.74 (Q-MW) vs. 4.50 (RR) for the bursty flow, and 2.74 (Q-MW) vs. 2.04 (RR) for the regular flow. This implies that Q-MW is not even a Pareto-optimal solution to minimizing the ages and we may be able to lower ages for all the flows.

IV. Age-Based MaxWeight Scheduling

In this section, we develop new policies that utilize a combination of *age* and *interarrival time* realizations/statistics in order to maintain fresh information at the receiver, instead of queue-lengths and delays. We apply a modified *drift-based heavy-traffic analysis* [28] to derive the heavy-traffic performance of our new policy in terms of the desired metric.

A. Algorithm Design

Under the assumption of heavy traffic loads, where $Q_l(t) > 1$ with high probability for all l with $\lambda_l > 0$, the evolution of the age (3) can be simplified as

$$U_l(t+1) = \begin{cases} U_l(t)+1, & \text{if } l \notin \mathbf{S}(t), \\ U_l(t)+1-X_{l,N_l(t)}, & \text{if } l \in \mathbf{S}(t), \end{cases} \quad (6)$$
$$\text{or} \quad U_l(t+1) = U_l(t)+1-X_{l,N_l(t)} \cdot S_l(t),$$

since $Q_l(t) > 1$ implies $U_l(t) \geq X_{l,N_l(t)}$. For any arrival $\boldsymbol{\lambda}$ strictly inside Λ, there is a stationary scheduler that schedules $\mathbf{S}^S(t)$ independent of the system state and satisfies, for small $\epsilon > 0$,

$$E[S_l^S(t)] \geq \lambda_l + \epsilon, \quad \text{for all } l. \quad (7)$$

We consider a Lyapunov function $V(t) := \frac{1}{2}\sum_l U_l(t)^2$. Let $\Delta V(t)$ denote the drift of the Lyapunov function. We have

$$\begin{aligned} &\Delta V(t) := E[V(t+1) - V(t) \mid \mathbf{U}(t) = \mathbf{U}] \\ &= \tfrac{1}{2}\textstyle\sum_l E\left[(U_l(t)+1-X_{l,N_l(t)} \cdot S_l(t))^2 - U_l(t)^2 | \mathbf{U}\right] \\ &\leq \textstyle\sum_l U_l \cdot E[1 - X_{l,N_l(t)} \cdot S_l^S(t) \mid \mathbf{U}] \\ &\quad + \textstyle\sum_l U_l \cdot E[X_{l,N_l(t)} \cdot (S_l^S(t) - S_l(t)) \mid \mathbf{U}] + C_1, \end{aligned} \quad (8)$$

where $C_1 := \frac{1}{2}|L|(1+X_{max}^2) \geq \frac{1}{2}\sum_l \left(1 - X_{l,N_l(t)} \cdot S(t)\right)^2$ is a constant. For the first term, since the interarrival time process and the service process of the stationary static scheduler are independent, and from (7), we can obtain that

$$U_l \cdot E\left[\left(1 - X_{l,N_l(t)} \cdot S_l^S(t)\right) \mid \mathbf{U}\right] \leq -\tfrac{\epsilon}{\lambda_l} U_l. \quad (9)$$

For the second term, we can minimize it by choosing $\mathbf{S}^I(t)$ as

$$\textbf{IA-MW:}\quad \mathbf{S}^I(t) = \underset{\mathbf{S}\in\mathcal{S}}{\operatorname{argmax}} \sum_{l\in L} U_l(t)\cdot X_{l,N_l(t)}\cdot S_l, \tag{10}$$

where $\mathcal{S}$ denotes the set of all feasible schedules. This immediately extends MaxWeight to take into account the product of Instantaneous interarrival time and Age[2] (thus denoted by IA-MW). As we will see later in Section V, however, the variation of the interarrival process often causes significant delaying of HoL packets. Hence, we also consider the class of scheduling policies that do not have the instantaneous interarrival times, in which case, we choose the schedule $\mathbf{S}^A(t)$ such that

$$\textbf{A-MW:}\quad \mathbf{S}^A(t) = \underset{\mathbf{S}\in\mathcal{S}}{\operatorname{argmax}} \sum_{l\in L} \frac{U_l(t)}{\lambda_l}\cdot S_l, \tag{11}$$

which takes into consideration average interarrival time and denoted by Age-based MaxWeight scheduling policy (A-MW).

Now, although A-MW may achieve good performance, it requires the information of arrival rate $\boldsymbol{\lambda}$, which may be unknown a priori. For more practical use, we can replace the arrival rate with measured value as

$$\textbf{mA-MW:}\quad \mathbf{S}^m(t) = \underset{\mathbf{S}\in\mathcal{S}}{\operatorname{argmax}} \sum_{l\in L} \frac{U_l(t)}{\hat{\lambda}_l(t)}\cdot S_l, \tag{12}$$

where $\hat{\lambda}_l(t) := \frac{1}{t}\sum_{\tau=0}^{t-1} A_l(\tau)$.

In the following, we focus on the performance characterization of A-MW due to mathematical tractability. Since $\hat{\lambda}_l(t)\to\lambda$ as $t\to\infty$, we claim that the performance of A-MW and mA-MW is close to each other, which will be verified through simluations in Section V.

B. Performance

In this section, We address the performance of A-MW in terms of stability, direction of state space collapse, and age bounds. We first show the stability of A-MW as follows.

Lemma 1: Age-based MaxWeight scheduling policy achieves the stability region Λ.

Proof: Under A-MW, the second term of (8) becomes

$$\begin{aligned}&\textstyle\sum_l U_l\cdot E\left[X_{l,N_l(t)}\cdot(S_l^S(t)-S_l^A(t))\mid \mathbf{U}\right]\\ &\quad= E\left[\textstyle\sum_l \frac{U_l}{\lambda_l}\cdot S_l^S(t) - \sum_l \frac{U_l}{\lambda_l}\cdot S_l^A(t)\ \middle|\mathbf{U}\right]\le 0,\end{aligned} \tag{13}$$

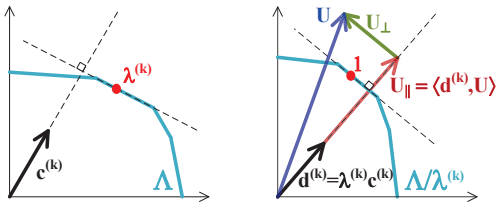
where the first equality comes from the independence between the schedules and $\mathbf{X}$. Combining (8), (9), and (13), we have

$$\Delta V(t)\le -\epsilon\textstyle\sum_{l\in L}\frac{U_l}{\lambda_l} + C_1, \tag{14}$$

which implies that i) A-MW has a negative Lyapunov drift for sufficiently large ages (thus achieves Λ), and ii) the information of the interarrival time instance is not required to achieve the (age) stability. ∎

Recall that the stability region of age is equivalent to the capacity region, which is shown in [28] to be bounded by K hyperplanes. Let $\mathcal{F}^{(k)}$ denote the k-th face of Λ, and let $\mathbf{c}^{(k)}$ denote the normal vector of $\mathcal{F}^{(k)}$ with $\|\mathbf{c}^{(k)}\|=1$. Then there is a constant $b^{(k)}$ such that

$$\langle \mathbf{c}^{(k)}, \mathbf{r}\rangle = b^{(k)} \text{ for all } \mathbf{r}\in\mathcal{F}^{(k)}. \tag{15}$$



(a) Original stability region (b) Scaled stability region

Fig. 5. Stability region and its scaled version (scaled by $1/\lambda^{(k)}$). Due to the componentwise division, the linearity is preserved. In the scaled stability region, we omit superscript (ϵ) or (ϵ,k) for the age vectors.

We define $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda := \{\frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda}^{(k)}}|\boldsymbol{\lambda}\in\Lambda\}$. All vector multiplications and divisions are componentwise. We consider $\boldsymbol{\lambda}^{(k)}\in$ relative interior of $\mathcal{F}^{(k)}$, and obtain $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda$ by scaling each element λ_l of $\boldsymbol{\lambda}$ in Λ with $\lambda_l^{(k)}$ as shown in Fig. 5. Due to the componentwise division, we have point-to-point mapping between Λ and $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda$, and the linearity is preserved. Hence, a face in Λ is mapped to a face in $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda$. Let $\mathcal{G}^{(k)}$ denote the face in $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda$ that corresponds to face $\mathcal{F}^{(k)}$ in Λ. We define $\mathbf{d}^{(k)} := \mathbf{c}^{(k)}\cdot\boldsymbol{\lambda}^{(k)}$, and given $\epsilon>0$, we choose an arrival vector $\boldsymbol{\lambda}^{(\epsilon)}$ such that

$$\frac{1}{\lambda_l^{(\epsilon)}} = \frac{1}{\lambda_l^{(k)}} + \epsilon\cdot\frac{\|\mathbf{d}^{(k)}\|}{d_l^{(k)}}, \tag{16}$$

for all l with non-zero $\lambda_l^{(k)}$ and $d_l^{(k)}$. We have $\lambda_l^{(\epsilon)}\le\lambda_l^{(k)},\forall l$.

Proposition 1 (State Space Collapse): Under the assumption of heavy traffic loads and independent interarrival times, the state space of the ages collapses under A-MW, to direction $\mathbf{d}^{(k)}$, as $\epsilon\to 0$.

To prove this, we basically follow the line of the analysis in [28]. However, the proof is not straightforward since the age processes do not evolve as the queue length processes: according to (6), they increase by 1 at each time slot, and decrease by the interarrival time. We scale the whole state space by $\boldsymbol{\lambda}^{(k)}$, and show that the mapping of the age to the hyperplane characterized by $\mathbf{d}^{(k)}$ approaches 0 as $\epsilon\to 0$. We refer the readers to Appendix A for the detailed proof.

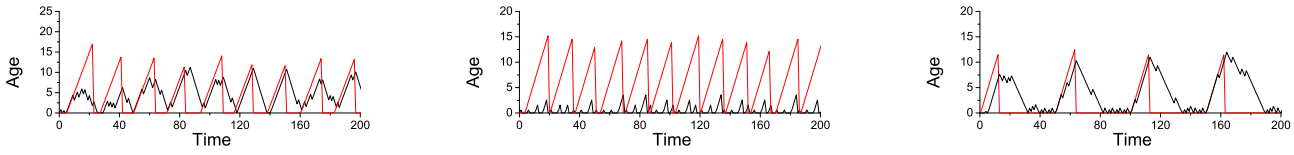
From the results in Appendix and Lemma 1 of [28], we can show that $\{\mathbf{U}(t)\}_t$ converges in distribution to a random variable $\overline{\mathbf{U}}$ with all bounded moments. For a vector $\mathbf{U}^{(\epsilon)}$, which is the age under A-MW with $\boldsymbol{\lambda}^{(\epsilon)}$, we define its parallel and perpendicular components with respect to $\mathbf{d}^{(k)}$ as follows:

$$\begin{aligned}\mathbf{U}_{\parallel}^{(\epsilon,k)} &:= \langle\tfrac{\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|},\mathbf{U}^{(\epsilon)}\rangle\tfrac{\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|},\\ \mathbf{U}_{\perp}^{(\epsilon,k)} &:= \mathbf{U}^{(\epsilon)} - \mathbf{U}_{\parallel}^{(\epsilon,k)}.\end{aligned} \tag{17}$$

Then, we have the following performance bounds, whose proofs can be found in Appendix B and C, respectively.

Proposition 2 (An Upper Bound): As $\epsilon\to 0$, A-MW

[2]We refer to [23] for intuitive explanation about the relationship between the interarrival times and the ages.



(a) $\lambda_{\text{regular}} = \lambda_{\text{bursty}}$ (b) $\lambda_{\text{regular}} = \lambda_{\text{bursty}}/3$ (c) $\lambda_{\text{regular}} = 3\lambda_{\text{bursty}}$

Fig. 6. Ages of two flows with different arrival rates under A-MW.

achieves that

$$\lim_{\epsilon \to 0} \epsilon E[\|\overline{\mathbf{U}}_{\|}\|] \leq \frac{1}{2} \cdot \langle (\frac{\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|})^2, (\boldsymbol{\sigma}^X)^2 \rangle, \tag{18}$$

where $\boldsymbol{\sigma}^X$ denotes the variance vector of the interarrival times. The following proposition shows that the performance bound under A-MW may not be tight.

Proposition 3 (A Lower Bound): For the class of scheduling policies that do not take into consideration the instantaneous interarrival times (i.e., interarrival-time-agnostic schedulers), the age performance is bounded by

$$\lim_{\epsilon \to 0} \epsilon E[\|\overline{\mathbf{U}}_{\|}\|] \geq \frac{1}{2} \langle (\mathbf{d}^{(k)})^2, (\boldsymbol{\sigma}^X)^2 \cdot (\boldsymbol{\lambda}^{(k)})^2 \rangle. \tag{19}$$

For the upper bound, we define $V_{\|}(\mathbf{U}, k) := \|\mathbf{U}_{\|}^{(\epsilon,k)}\|^2$ and note that its drift $E[\Delta V_{\|}(\mathbf{U}, k)]$ is zero from the age stability under A-MW. Starting from the zero drift, we carefully derive the equations in terms of $E[\|\mathbf{U}_{\|}^{(\epsilon,k)}\|]$ when $\epsilon \to 0$, which results in the upper bound. Technical difficulties mainly come from the product form of the processes $\mathbf{X} \cdot \mathbf{S}$ and the non-linear relationship between $\boldsymbol{\lambda}^{(\epsilon)}$ and $\boldsymbol{\lambda}^{(k)}$. For the lower bound, we consider a single-queue server with constant arrival $\langle \mathbf{d}^{(k)}, \mathbf{1} \rangle$ and departure $\max_{\mathbf{S} \in \Lambda} \langle \mathbf{d}^{(k)}, \mathbf{X} \cdot \mathbf{S} \rangle$, which outperforms A-MW in terms of age. The product form of the processes $\mathbf{X} \cdot \mathbf{S}$ again becomes the technical difficulty. We restrict our attention to the class of interarrival-time-agnostic schedulers, and show that the departure process of the single-queue server that achieves $\boldsymbol{\lambda}^{(k)}$ is an optimal solution, which leads to (19).

V. Numerical Results

In this section, we evaluate the performance of A-MW. We first present the behavior of A-MW with two flows (one bursty and one regular traffic), and then compare the performance of A-MW with those of RR, Q-MW, and D-MW. Finally, we observe the state space collapse of the ages under A-MW.

Under the same scenarios as in Section III, we can observe the age performance of A-MW under equal and unequal packet arrivals as in Fig. 6. See Figs. 3 and 4 for comparison with RR, Q-MW, and D-MW. The age performance of A-MW is similar to that of D-MW, which can be also observed by the total average ages: 9.73 when $\lambda_{\text{regular}} = \lambda_{\text{bursty}}$, 6.86 when $\lambda_{\text{regular}} = \lambda_{\text{bursty}}/3$, and 5.65 when $\lambda_{\text{regular}} = 3\lambda_{\text{bursty}}$.

Next, we further evaluate the performance of A-MW in terms of queue lengths, packet delays, and normalized age. Besides RR, Q-MW, and D-MW, we also consider IA-MW of (10) that takes into account instantaneous interarrival times, and mA-MW of (12). We consider a simple network scenario with one base station and 4 users (flows) as shown in Fig. 7. The base

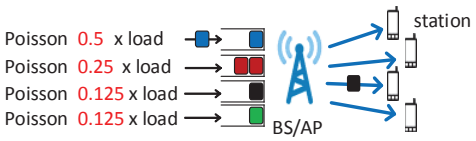
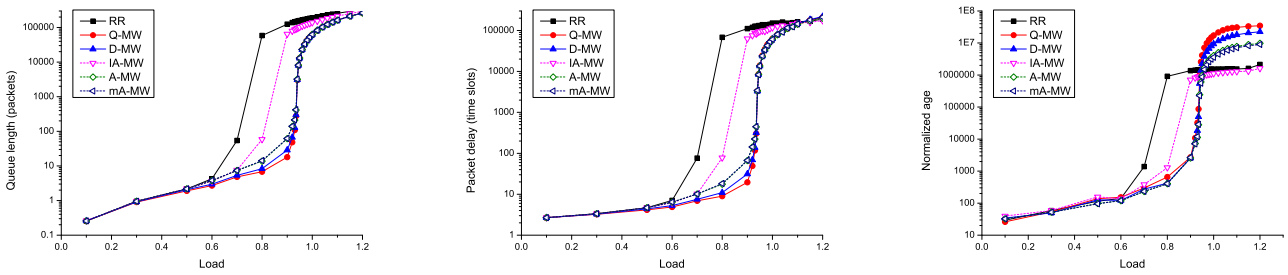


Fig. 7. Network topology with 4 flows.

station has 4 downlinks, where each link is dedicated to a flow. Packets for each flow arrive at the base station, stored in separate per-flow queues, and served through the links. At a given time slot, the channel of each link is either *on* or *off* with probability 0.5, and the scheduler of the base station can choose one link with on channel. Once a link is chosen, it can serve one packet during the time slot. Packets for each flow arrive following a Poisson process with mean arrival rate $\boldsymbol{\lambda} = \rho \cdot \{0.5, 0.25, 0.125, 0.125\}$, where load ρ is the scaling factor of the arrival rate vector. We simulate the system for 10^6 time slots under different traffic loads. We use 10 different random seeds, and in each simulation run, we measure moving average of total queue lengths $\sum_l Q_l(t)$, total packet delays $\frac{1}{\sum_l N_l(t)} \sum_{l,i} (t'_{l,i} - t_{l,i})$, and total normalized ages $\sum_l \frac{U_l(t)}{\lambda_l}$.

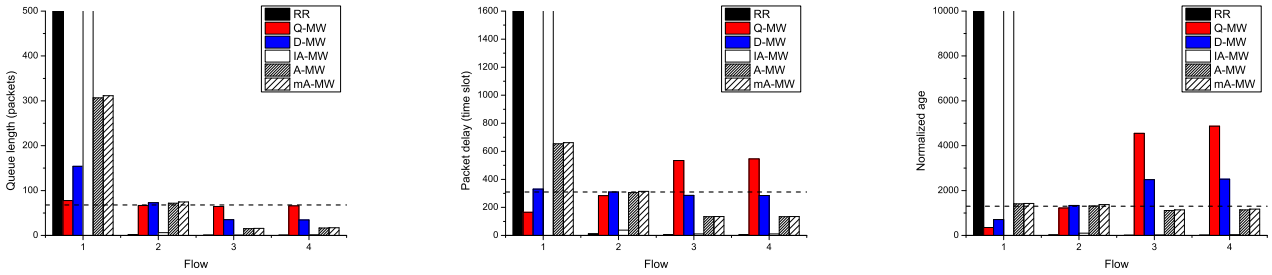
Fig. 8 show, in log scale, the measured values after the simulations end, and each point represents an average over the 10 simulation runs. Given our setting, $\rho = 1 - (0.5)^4 = 0.9375$ is the boundary of Λ. First, we can observe that under RR and IA-MW, the queue lengths, the packet delays, and the ages start soaring before the load increases close enough to the boundary, which implies that they may not achieve the stability region. For IA-MW, the variance of interarrival times seems to often cause excessive delay for the packets with short interarrival times, which degrades the performance. Second, the performance of A-MW and mA-MW are very similar. By replacing the arrival rate λ_l with our measurement $\frac{1}{t} \sum_{\tau=0}^{t} A_l(\tau)$, we can implement the scheduling policy without the rate information of the flows. Third, in the queue lengths and the packet delays, Q-MW and D-MW outperform A-MW and mA-MW, but the differences reduce as the load approaches the boundary. In contrast, A-MW and mA-MW outperform Q-MW and D-MW in the normalized ages, and there are substantial differences remain at the boundary and even the beyond. This shows that A-MW and mA-MW achieve higher age performance at no significant cost of queue length and delay.

Fig. 9 provides the per-flow performance when $\rho = 0.935$, which is more than 99.7% of the capacity. It clarifies the differences of Q-MW, D-MW, and A-MW in the per-flow delay performance. Flows are numbered in the decreasing order of



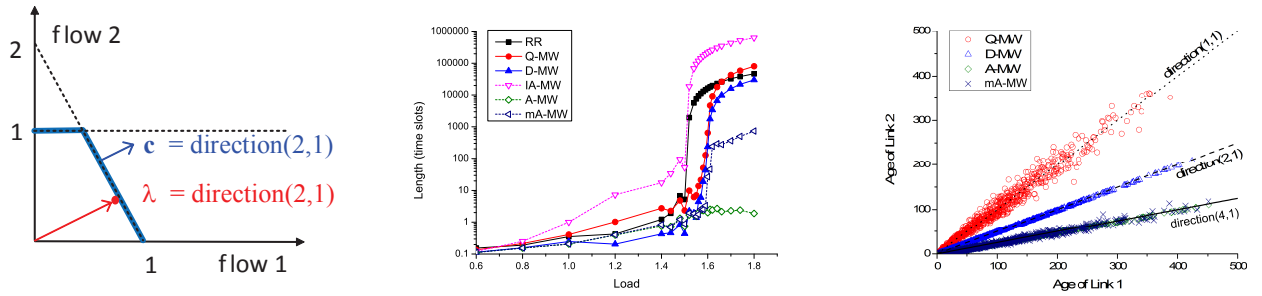
(a) Queue lengths (b) Packet delays (c) Normalized ages

Fig. 8. Performance with different traffic loads.



(a) Queue lengths (b) Packet delays (c) Normalized ages

Fig. 9. Per-flow performance when load $\rho = 0.935$.



(a) Stability region Λ (b) $\|\mathbf{U}_\perp\|$ with different loads (c) Ages of each flow (load $\rho = 1.59$)

Fig. 10. State space collapse.

the arrival rate. In comparison of the queue lengths shown in Fig. 9(a), Q-MW achieves almost equal queue length (as denoted by the dotted line) over all the flows. In packet delays, D-MW achieves equal per-packet delays over the flows in Fig. 9(b). Finally, Fig. 9(c) shows that A-MW and mA-MW achieve equal normalized age over the flows (as denoted by the dotted line). Considering the Little's law that associates the queue lengths (that Q-MW schedules with as in (4)) and the packet delays (that D-MW schedules with as in (5)) by the arrival rate as $\frac{1}{\lambda_l}$, one may expect that the performances under Q-MW and D-MW are also related by $\frac{1}{\lambda_l}$ as shown in the results[3]. A similar relationship that can be expected between D-MW (5) and A-MW (11) is supported by our results. We emphasize that the property of A-MW that gives a priority to the flow with a small arrival rate is desirable. A traffic source can decrease its transmission rate to achieve better delay performance, which will improve the overall delay performance by decreasing the traffic load.

Finally, we investigate the state space collapse. We consider a network with two users. The network settings are the same, except that the channel is *on* with probability 1 and 0.5 for user 1 and user 2, respectively, and the link for user 2 can serve up to 2 packets if it is scheduled. For the non-unit service rate, we have modified Q-MW (4) $\mathbf{S}^Q(t) = \text{argmax}_{\mathbf{S}\in\mathcal{S}} \sum_{l\in L} Q_l(t) \cdot S_l \cdot r_l$, where r_l denote the service rate of link l. The other policies of D-MW, IA-MW, A-MW, and mA-MW are also modified accordingly. In this scenario, the stability region is as shown in Fig. 10(a). Consider $\boldsymbol{\lambda} = \rho \cdot \{0.5, 0.25\}$. Then, the slope is the face and we have the normal vector $\mathbf{c}^{(k)} = \frac{1}{\sqrt{5}}\{2, 1\}$, and thus

[3]We note that the packet delays are a per-packet average while the queue lengths and the ages are a time average. However, our statement will hold under Poisson arrival processes due to PASTA.

$\mathbf{d}^{(k)} = \mathbf{c}^{(k)} \cdot \boldsymbol{\lambda} = \frac{\rho}{\sqrt{5}} \cdot \{1, 0.25\}$. Note that the arrival rate is on the boundary of Λ when $\rho = 1.6$. Fig. 10(b) demonstrates that as ρ increases, the perpendicular element $\|\mathbf{U}_\perp\|$ of the age keeps increasing under RR, Q-MW, D-MW, and IA-MW. In contrast, A-MW achieves a finite $\|\mathbf{U}_\perp\|$, which verifies the state space collapse to direction $\mathbf{d}^{(k)}$ under A-MW. mA-MW has slowly increasing $\|\mathbf{U}_\perp\|$ due to some measurement errors, but it has much smaller $\|\mathbf{U}_\perp\|$ than RR, Q-MW, and D-MW. Fig. 10(c) directly shows the evolution of the ages for the flows $\{U_1(t), U_2(t)\}$ when $\rho = 1.59$. Since D-MW tries to have $U_1(t) = 2U_2(t)$, where the doubling is due to the high link rate, it has the ages to the direction of $\{2, 1\}$. Q-MW tries to have $Q_1(t) = 2Q_2(t)$, hence, through Little's law, achieves $E[D_1(t)] = E[\frac{Q_1(t)}{0.5\rho}] = E[\frac{2Q_2(t)}{0.5\rho}] = E[D_2(t)]$, i.e., has direction $\{1, 1\}$. A-MW and mA-MW tries to have $U_1(t) \cdot 2 = 2U_2(t) \cdot 4$, and thus the ages evolve along direction $\{4, 1\}$.

While we have shown that A-MW and mA-MW achieve good performance and desirable properties of state-space collapse under heavy traffic loads, the schemes can be considered as a weighted version of D-MW, and needs a long-term averaging of interarrival times, which may make the scheme less responsive to traffic changes. To this end, a time-weighted moving average of the interarrival times could be helpful. Taking into consideration the low performance of IA-MW (i.e., without averaing interarrival times), finding a good factor of time-averaging would be an interesting open problem.

VI. Concluding Remarks

In this work, we address the scheduling problem in wireless networks with a focus on the information freshness and the delay alignment, which are of great importance to the systems where the effectiveness of the control decisions depends critically on the delay and synchronism of the system state information. We start with inefficiency of conventional approaches in maintaining fresh information updates of multiple continuous flows, and show the critical value of both age and interarrival times. We develop new schedulers, with and without the knowledge of arrival rates, that account for both age information and interarrival times of incoming packets, and characterize its performance under heavy-traffic condition. To elaborate, we show that it achieves the state space collapse in a properly scaled coordination system, and provide its upper and lower performance bounds. Although the analytical results are obtained under heavy-traffic conditions, we observe through numerical results that the scheduler achieves desirable freshness performance even in lightly-loaded conditions. In addition, the scheduler has good long-term performance in throughput and average delay, while also maintaining equally-up-to-dated information from multiple sources.

References

[1] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource Allocation and Cross-Layer Control in Wireless Networks," *Found. Trends Netw.*, vol. 1, no. 1, pp. 1–144, 2006.

[2] X. Lin, N. B. Shroff, and R. Srikant, "A Tutorial on Cross-Layer Optimization in Wireless Networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, August 2006.

[3] M. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.

[4] I.-H. Hou, V. Borkar, P. Kumar, *et al.*, *A Theory of QoS for Wireless*. IEEE, 2009.

[5] J. Liu, Y. Yi, A. Proutiere, M. Chiang, and H. V. Poor, "Towards utility-optimal random access without message passing," *Wireless Communications and Mobile Computing*, vol. 10, no. 1, pp. 115–128, 2010.

[6] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press, 2013.

[7] L. Tassiulas and A. Ephremides, "Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximal Throughput in Multihop Radio Networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, December 1992.

[8] M. Andrews, K. Kumaran, K.Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Providing Quality of Service over a Shared Wireless Link," *IEEE Communications Magazine*, February 2001.

[9] S. Shakkottai, R. Srikant, and A. Stolyar, "Pathwise optimality of the exponential scheduling rule for wireless channels," *Advances in Applied Probability*, vol. 36, pp. 1021–1045, 2004.

[10] A. L. Stolyar, "On the Asymptotic Optimality of the Gradient Scheduling Algorithm for Multiuser Throughput Allocation," *Oper. Res.*, vol. 53, no. 1, pp. 12–25, 2005.

[11] A. Eryilmaz, R. Srikant, and J. R. Perkins, "Stable Scheduling Policies for Fading Wireless Channels," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 411–424, 2005.

[12] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic Power Allocation and Routing for Time-varying Wireless Networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, 2005.

[13] M. J. Neely, E. Modiano, and L. Chih-Ping, "Fairness and Optimal Stochastic Control for Heterogeneous Networks," in *IEEE INFOCOM*, March 2005, pp. 1723–1734.

[14] A. L. Stolyar and K. Ramanan, "Largest Weighted Delay First Scheduling: Large Deviations and Optimality," *Ann. Appl. Probab.*, vol. 11, no. 1, pp. 1–48, February 2001.

[15] B. Sadiq and G. de Veciana, "Throughput Optimality of Delay-driven MaxWeight Scheduler for a Wireless System with Flow Dynamics," in *Proc. Allerton Conference on Communications, Control and Computing*, October 2009.

[16] M. Neely, "Delay-Based Network Utility Maximization." *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 41–54, 2013.

[17] B. Ji, C. Joo, and N. B. Shroff, "Delay-Based Back-Pressure Scheduling in Multi-Hop Wireless Networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1539–1552, October 2013.

[18] B. Li, A. Eryilmaz, and R. Srikant, "On the Universality of Age-based Scheduling in Wireless Networks," in *IEEE INFOCOM*, April 2015.

[19] I. Hou and P. R. Kumar, "Scheduling Heterogeneous Real-Time Traffic over Fading Wireless Channels," in *IEEE INFOCOM*, March 2010.

[20] R. Urgaonkar and M. Neely, "Opportunistic Scheduling with Reliability Guarantees in Cognitive Radio Networks," in *IEEE INFOCOM*, 2008.

[21] R. Li, A. Eryilmaz, and B. Li, "Throughput-Optimal Scheduling with Regulated Inter-Service Times," in *IEEE INFOCOM*, April 2013.

[22] B. Li, R. Li, and A. Eryilmaz, "Heavy-Traffic-Optimal Scheduling Design with Regular Service Guarantees in Wireless Networks," in *ACM MobiHoc*, July 2013.

[23] S. Kau, R. Yates, and M. Gruteser, "Real-Time Status: How Often Should One Update?" in *IEEE INFOCOM*, 2012.

[24] R. D. Yates and S. Kaul, "Real-Time Status Updating: Multiple Sources," in *ISIT*, July 2012.

[25] R. D. Yates, "Lazy is Timely: Status Updates by an Energy Harvesting Source," in *ISIT*, June 2015.

[26] Y. Sun, E. Uysal-Biyikoglu, R. Yates, C. E. Koksal, , and N. B. Shroff, "Update or Wait: How to Keep Your Data Fresh," in *IEEE INFOCOM*, April 2016.

[27] Q. He, D. Yuan, and A. Ephremides, "Optimizing freshness of information: On minimum age link scheduling in wireless systems," in *WiOpt*, May 2016.

[28] A. Eryilmaz and R. Srikant, "Asymptotically Tight Steady-state Queue Length Bounds Implied by Drift Conditions," *Queueing Systems*, vol. 72, no. 3-4, pp. 311–359, 2012.

[29] B. Hajek, "Hitting-time and Occupaton-time Bounds implied by Drift Analysis with Applications," *Advances in Applied Probability*, vol. 14, no. 3, pp. 502–525, September 1982.

[30] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

APPENDIX

A. State space collapse

Recall that $\mathcal{G}^{(k)}$ denote the face in $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda$ that corresponds to face $\mathcal{F}^{(k)}$ after the componentwise mapping of Λ. Let $\mathcal{H}^{(k)}$ denote the hyperplane that characterizes $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda$ and includes face $\mathcal{G}^{(k)}$. From $\mathbf{d}^{(k)} = \mathbf{c}^{(k)} \cdot \boldsymbol{\lambda}^{(k)}$, we have $\langle \mathbf{d}^{(k)}, \mathbf{q}\rangle = b^{(k)}$, for all $\mathbf{q} \in \mathcal{G}^{(k)}$, since for any $\mathbf{q} \in \mathcal{G}^{(k)}$, we can find $\mathbf{r} \in \mathcal{F}^{(k)}$ such that $\mathbf{r} = \mathbf{q} \cdot \boldsymbol{\lambda}^{(k)}$ due to the mapping, and then

$$\langle \mathbf{d}^{(k)}, \mathbf{q}\rangle = \langle \mathbf{c}^{(k)} \cdot \boldsymbol{\lambda}^{(k)}, \mathbf{r}/\boldsymbol{\lambda}^{(k)}\rangle = \langle \mathbf{c}^{(k)}, \mathbf{r}\rangle = b^{(k)}, \quad (20)$$

where the last equality from (15). This means that $\mathbf{d}^{(k)}$ is normal[4] to $\mathcal{G}^{(k)}$ (while $\|\mathbf{d}^{(k)}\| \neq 1$.)

Let $\mathbf{1} = \{1, 1, \ldots, 1\}$. From $\boldsymbol{\lambda}^{(k)} \in \mathcal{F}^{(k)}$, we have $\mathbf{1} \in \mathcal{G}^{(k)}$. Further, from $\|\mathbf{c}^{(k)}\| = 1$, we have

$$\begin{aligned} \langle \mathbf{d}^{(k)}, \mathbf{1}\rangle &= \langle \mathbf{c}^{(k)}, \boldsymbol{\lambda}^{(k)}\rangle = b^{(k)}, & (21)\\ \langle \mathbf{d}^{(k)}, \boldsymbol{\lambda}^{(\epsilon)}/\boldsymbol{\lambda}^{(k)}\rangle &= b^{(k)} - \epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{1}, \boldsymbol{\lambda}^{(\epsilon)}\rangle, & (22)\\ \|\mathbf{d}^{(k)}\| &\leq \|\mathbf{c}^{(k)}\| \cdot \|\boldsymbol{\lambda}^{(k)}\| = \|\boldsymbol{\lambda}^{(k)}\|. & (23) \end{aligned}$$

where the arrival vector $\boldsymbol{\lambda}^{(\epsilon)}$ is specified in (16). From our selection $\boldsymbol{\lambda}^{(\epsilon)}$, we always have $\lambda_l^{(\epsilon)} \leq \lambda_l^{(k)}$ for all l. Let $\mathbf{U}^{(\epsilon)}(t)$ denote the ages under $\boldsymbol{\lambda}^{(\epsilon)}$.

For a vector $\mathbf{U}^{(\epsilon)}(t)$, we define its parallel and perpendicular components with respect to $\mathbf{d}^{(k)}$ as follows:

$$\begin{aligned} \mathbf{U}_{\|}^{(\epsilon,k)}(t) &:= \langle \tfrac{\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|}, \mathbf{U}^{(\epsilon)}(t)\rangle \tfrac{\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|},\\ \mathbf{U}_{\perp}^{(\epsilon,k)}(t) &:= \mathbf{U}^{(\epsilon)} - \mathbf{U}_{\|}^{(\epsilon,k)}(t). \end{aligned} \quad (24)$$

In the sequel, we omit scripts t and ϵ for simplicity, and denote $\mathbf{U}^{(\epsilon)}(t), \mathbf{U}_{\|}^{(\epsilon,k)}(t), \mathbf{U}_{\perp}^{(\epsilon,k)}(t)$ by $\mathbf{U}, \mathbf{U}_{\|}^{(k)}, \mathbf{U}_{\perp}^{(k)}$, respectively.

Let us consider the following three Lyapunov functions. $V(\mathbf{U}) := \|\mathbf{U}\|^2$, $V_{\perp}(\mathbf{U}, k) := \|\mathbf{U}_{\perp}^{(k)}\|$, and $V_{\|}(\mathbf{U}, k) := \|\mathbf{U}_{\|}^{(k)}\|^2$. Given $\boldsymbol{\lambda}^{(k)} \in \mathcal{F}^{(k)}$ (relatively inside) and $\delta > 0$, we define

$$B_{\delta}^{(k)} := \{\mathbf{r} \mid \|\mathbf{1} - \tfrac{\mathbf{r}}{\boldsymbol{\lambda}^{(k)}}\| \leq \delta \text{ and } \tfrac{\mathbf{r}}{\boldsymbol{\lambda}^{(k)}} \in \mathcal{H}^{(k)}\}. \quad (25)$$

Since $\boldsymbol{\lambda}^{(k)}$ is relatively inside $\mathcal{F}^{(k)}$, $\boldsymbol{\lambda}^{(k)}/\boldsymbol{\lambda}^{(k)} = \mathbf{1}$ is relatively inside $\mathcal{G}^{(k)}$, and thus there exists a small $\delta^{(k)} > 0$ such that $\frac{1}{\boldsymbol{\lambda}^{(k)}} B_{\delta^{(k)}}^{(k)}$ lies strictly inside $\mathcal{G}^{(k)}$. The following lemma implies the state space collapse as $\epsilon \to 0$.

Lemma 2:

$$E[\Delta V_{\perp}(\mathbf{U}, k)|\mathbf{U}] \leq -\delta^{(k)} + \epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \|\tfrac{1}{\mathbf{d}^{(k)}}\| + \tfrac{C_1}{\|\mathbf{U}_{\perp}^{(k)}\|}. \quad (26)$$

Lemma 2 implies the existence of constants $\{N_r^{(k)}\}$ such that $E[\|\overline{\mathbf{U}}_{\perp}^{(\epsilon,k)}\|^r] \leq N_r^{(k)}$ for sufficiently small ϵ and each $r = 1, 2, \ldots$ in our system. (See [29] or Lemma 1 of [28].)

[4]For any $\mathbf{q}_1, \mathbf{q}_2 \in \mathcal{G}^{(k)}$, we can find some $\mathbf{r}_1$ and $\mathbf{r}_2$ such that $\mathbf{r}_1 = \mathbf{q}_1 \cdot \boldsymbol{\lambda}^{(k)}$ and $\mathbf{r}_2 = \mathbf{q}_2 \cdot \boldsymbol{\lambda}^{(k)}$. Then $\langle \mathbf{d}^{(k)}, \mathbf{q}_1 - \mathbf{q}_2\rangle = \langle \mathbf{c}^{(k)} \cdot \boldsymbol{\lambda}^{(k)}, \mathbf{r}_1/\boldsymbol{\lambda}^{(k)} - \mathbf{r}_2/\boldsymbol{\lambda}^{(k)}\rangle = \langle \mathbf{c}^{(k)}, \mathbf{r}_1 - \mathbf{r}_2\rangle$. From the construction of $\mathcal{G}^{(k)}$, $\mathbf{r}_1$ and $\mathbf{r}_2$ are two points on $\mathcal{F}^{(k)}$ and we have $\langle \mathbf{c}^{(k)}, \mathbf{r}_1 - \mathbf{r}_2\rangle = 0$, since $\mathbf{c}^{(k)}$ is normal to $\mathcal{F}^{(k)}$.

Proof: We start from the following equation shown in [28].

$$\begin{aligned} &E[\Delta V_{\perp}(\mathbf{U}, k)|\mathbf{U}]\\ &\leq E\left[\tfrac{1}{2\|\mathbf{U}_{\perp}^{(k)}\|} \cdot \left(\Delta V(\mathbf{U}) - \Delta V_{\|}(\mathbf{U}, k)\right)|\mathbf{U}\right]. \end{aligned} \quad (27)$$

We consider each drift $E[\Delta V(\mathbf{U})|\mathbf{U}]$ and $E[\Delta V_{\|}(\mathbf{U})|\mathbf{U}]$ one by one.

For $E[\Delta V(\mathbf{U})|\mathbf{U}]$, we have,

$$\begin{aligned} &E[\Delta V(\mathbf{U})|\mathbf{U}]\\ &= E[\ \|\mathbf{U}(t+1)\|^2 - \|\mathbf{U}(t)\|^2\ |\mathbf{U}(t) = \mathbf{U}\ ] \quad (28)\\ &\leq 2E[\langle \mathbf{U}, \mathbf{1} - \mathbf{X} \cdot \mathbf{S}^A\rangle|\mathbf{U}] + 2C_1, \end{aligned}$$

where $C_1 := \frac{1}{2} \cdot L \cdot (1 + X_{max})^2$, since $\mathbf{U}(t+1) = \mathbf{U}(t) + \mathbf{1} - \mathbf{X}(t) \cdot \mathbf{S}^A(t)$ under heavy traffic. From the independence of $\mathbf{X}$, the first term can be drived as

$$\begin{aligned} &2E[\langle \mathbf{U}, \mathbf{1} - \mathbf{X} \cdot \mathbf{S}^A\rangle|\mathbf{U}] = 2\langle \mathbf{U}, \mathbf{1} - E[\mathbf{S}^A|\mathbf{U}]/\boldsymbol{\lambda}^{(\epsilon)}\rangle\\ &= 2\langle \mathbf{U}, \mathbf{1} - E[\mathbf{S}^A|\mathbf{U}]/\boldsymbol{\lambda}^{(k)}\rangle - 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle\\ &\overset{(a)}{\leq} 2 \min_{\mathbf{r} \in B_{\delta^{(k)}}^{(k)}} \langle \mathbf{U}, \mathbf{1} - \mathbf{r}/\boldsymbol{\lambda}^{(k)}\rangle - 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle\\ &\overset{(b)}{=} 2 \min_{\mathbf{r} \in B_{\delta^{(k)}}^{(k)}} \langle \mathbf{U}_{\perp}^{(k)}, \mathbf{1} - \mathbf{r}/\boldsymbol{\lambda}^{(k)}\rangle\\ &\qquad - 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}_{\perp}^{(k)} + \mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle\\ &\overset{(c)}{=} -2\delta^{(k)} \cdot \|\mathbf{U}_{\perp}^{(k)}\| - 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}_{\perp}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle\\ &\qquad - 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle, \end{aligned} \quad (29)$$

where (a) comes from the scheduling of A-MW that maximizes $\langle \mathbf{U}, E[\mathbf{S}^A|\mathbf{U}]/\boldsymbol{\lambda}^{(\epsilon)}\rangle$, (b) holds since $(\mathbf{1} - \mathbf{r}/\boldsymbol{\lambda}^{(k)}) \in \mathcal{H}^{(k)}$ and it is perpendicular to $\mathbf{U}_{\|}^{(k)}$, and (c) holds since the minimum will be obtained by choosing the point in $B_{\delta^{(k)}}^{(k)}$ to the opposite direction of $\mathbf{U}_{\perp}^{(k)}$. Further, the last two tersm of the last equation in (29) can be rewritten as

$$\begin{aligned} \langle \mathbf{U}_{\perp}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle &= \|\mathbf{U}_{\perp}^{(k)}\| \cdot \|E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\| \cdot \cos\theta_1,\\ \langle \mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle &= \|\mathbf{U}_{\|}^{(k)}\| \cdot \|E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\| \cdot \cos\theta_2, \end{aligned}$$

where $\theta_1 := \angle(\mathbf{U}_{\perp}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)})$ and $\theta_2 := \angle(\mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)})$.

For $E[\Delta V_{\|}(\mathbf{U})|\mathbf{U}]$, from

$$\|\mathbf{U}_{\|}^{(k)}\| = \langle \tfrac{\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|}, \mathbf{U}\rangle \tfrac{\|\mathbf{d}^{(k)}\|}{\|\mathbf{d}^{(k)}\|} = \tfrac{1}{\|\mathbf{d}^{(k)}\|}\langle \mathbf{d}^{(k)}, \mathbf{U}\rangle, \quad (30)$$

we have

$$\begin{aligned} &E[\Delta V_{\|}(\mathbf{U})|\mathbf{U}]\\ &= E[\ \|\mathbf{U}_{\|}^{(k)}(t+1)\|^2 - \|\mathbf{U}_{\|}^{(k)}(t)\|^2\ |\mathbf{U}(t) = \mathbf{U}]\\ &= \frac{1}{\|\mathbf{d}^{(k)}\|^2} \cdot E[\langle \mathbf{d}^{(k)}, \mathbf{U} + \mathbf{1} - \mathbf{X} \cdot \mathbf{S}^A\rangle^2 - \langle \mathbf{d}^{(k)}, \mathbf{U}\rangle^2\ |\mathbf{U}]\\ &\geq \frac{2}{\|\mathbf{d}^{(k)}\|^2} \cdot \langle \mathbf{d}^{(k)}, \mathbf{U}\rangle \cdot \langle \mathbf{d}^{(k)}, \mathbf{1} - E[\mathbf{S}^A|\mathbf{U}]/\boldsymbol{\lambda}^{(\epsilon)}\rangle. \end{aligned}$$

From (16) and (30), we have

$$\begin{aligned}
&\langle \mathbf{d}^{(k)}, \mathbf{U}\rangle \cdot \langle \mathbf{d}^{(k)}, \mathbf{1} - E[\mathbf{S}^A|\mathbf{U}]/\boldsymbol{\lambda}^{(\epsilon)}\rangle \\
&\quad = \|\mathbf{d}^{(k)}\| \cdot \|\mathbf{U}_{\|}^{(k)}\| \cdot \Big(\langle \mathbf{d}^{(k)}, \mathbf{1} - E[\mathbf{S}^A|\mathbf{U}]/\boldsymbol{\lambda}^{(k)}\rangle \\
&\qquad -\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{d}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle\Big) \\
&\quad \geq -\epsilon \cdot \|\mathbf{d}^{(k)}\|^3 \cdot \|\mathbf{U}_{\|}^{(k)}\| \cdot \|E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\| \cdot \cos\theta_3,
\end{aligned}$$

where $\theta_3 := \angle(\mathbf{d}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)})$. The inequality holds since $\langle \mathbf{d}^{(k)}, \mathbf{1}\rangle = b^{(k)}$ and $\langle \mathbf{d}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\boldsymbol{\lambda}^{(k)}\rangle \leq b^{(k)}$ for any feasible schedules.

Note that $\theta_2 = \angle(\mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)})$ equals to $\angle(\mathbf{d}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}) = \theta_3$. Thus, we can obtain that

$$\begin{aligned}
&E[\Delta V_{\|}(\mathbf{U})|\mathbf{U}] \\
&\quad \geq -2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \|\mathbf{U}_{\|}^{(k)}\| \cdot \|E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\| \cdot \cos\theta_2 \\
&\quad = -2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle.
\end{aligned} \tag{31}$$

From (27), (28), (29), and (31), we can obtain

$$\begin{aligned}
&E[\Delta V_{\perp}(\mathbf{U})|\mathbf{U}] \\
&\leq \tfrac{1}{2\|\mathbf{U}_{\perp}^{(k)}\|}\Big(-2\delta^{(k)} \cdot \|\mathbf{U}_{\perp}^{(k)}\| \\
&\quad - 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \|\mathbf{U}_{\perp}^{(k)}\| \cdot \|E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\| \cdot \cos\theta_1 \\
&\quad - 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle + 2C_1 \\
&\quad + 2\epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{U}_{\|}^{(k)}, E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\rangle\Big) \\
&\leq -\delta^{(k)} + \epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \|\tfrac{\mathbf{1}}{\mathbf{d}^{(k)}}\| + \tfrac{C_1}{\|\mathbf{U}_{\perp}^{(k)}\|},
\end{aligned} \tag{32}$$

where the last inequality comes from the fact that $|\cos\theta_1| \leq 1$, and that $0 \leq E[S_l^A|\mathbf{U}] \leq 1$ for all l, which implies $\|E[\mathbf{S}^A|\mathbf{U}]/\mathbf{d}^{(k)}\| \leq \|\frac{\mathbf{1}}{\mathbf{d}^{(k)}}\|$. ■

Hence, we obtain that $E[\Delta V_{\perp}(\mathbf{U})|\mathbf{U}] < 0$, with sufficiently small ϵ and sufficiently large $\|\mathbf{U}_{\perp}^{(k)}\|$, which implies that the state space collapses to the direction of $\mathbf{d}^{(k)}$ as $\epsilon \to 0$.

B. An upper bound

From the results in Section A and Lemma 1 of [28], $\{\mathbf{U}(t)\}_t$ converges in distribution to a random variable $\overline{\mathbf{U}}$ with all bounded moments, i.e., $E[\|\overline{\mathbf{U}}\|^r] \leq \infty$ for each $r = 1, 2, \cdots$. Let $\mathbf{S}^A(\overline{\mathbf{U}})$ denote a random variable for given $\overline{\mathbf{U}}$ that represents the scheduling vector chosen by A-MW, where the randomness comes from the interarrival times.

From the stability result of Lemma 1 and the fact that $\|\overline{\mathbf{U}}_{\|}\|^2 \leq \|\overline{\mathbf{U}}\|^2$, we have $E[\Delta V_{\|}(\overline{\mathbf{U}}, k)] = 0$, which results in the following lemma.

Lemma 3: For any positive vector $\mathbf{d}^{(k)}$, in steady state, we have

$$\begin{aligned}
&2E[\langle \mathbf{d}^{(k)}, \overline{\mathbf{U}}\rangle \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(\epsilon)} - \mathbf{1}\rangle] \\
&\qquad = E[\langle \mathbf{d}^{(k)}, \mathbf{1} - \mathbf{X} \cdot \mathbf{S}^A(\overline{\mathbf{U}})\rangle^2].
\end{aligned} \tag{33}$$

Proof: From the drift of the $V_{\|}(\mathbf{U}, k)$,

$$\begin{aligned}
&E[\Delta V_{\|}(\mathbf{U}, k)] \\
&= \tfrac{1}{\|\mathbf{d}^{(k)}\|^2} \cdot E[\langle \mathbf{d}^{(k)}, \mathbf{U}(t+1)\rangle^2 - \langle \mathbf{d}^{(k)}, \mathbf{U}(t)\rangle^2 | \mathbf{U}(t) = \mathbf{U}] \\
&= \tfrac{1}{\|\mathbf{d}^{(k)}\|^2} \cdot \Big(E[2\langle \mathbf{d}^{(k)}, \overline{\mathbf{U}}\rangle \langle \mathbf{d}^{(k)}, \mathbf{1} - \mathbf{S}^A/\boldsymbol{\lambda}^{(\epsilon)}\rangle] \\
&\qquad + E[\langle \mathbf{d}^{(k)}, \mathbf{1} - \mathbf{X} \cdot \mathbf{S}^A\rangle^2]\Big),
\end{aligned}$$

due to the independence of $\mathbf{X}$. From $E[\Delta V_{\|}(\overline{\mathbf{U}}, k)] = 0$, the result follows. ■

We now focus on the both sides of (33).

For a bound on $E[\langle \mathbf{d}^{(k)}, \overline{\mathbf{U}}\rangle \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(\epsilon)} - \mathbf{1}\rangle])$, we first note that the geometry of the capacity region Λ has a finite number of faces. Then, for each face $\mathcal{G}^{(k)}$ of $\frac{1}{\boldsymbol{\lambda}^{(k)}}\Lambda$, there exists $\theta^{(k)} \in (0, \pi/2]$ such that

$$\langle \mathbf{d}^{(k)}, \mathbf{S}^A(\mathbf{U})/\boldsymbol{\lambda}^{(k)}\rangle = b^{(k)}, \text{ for all } \mathbf{U} \text{ with } \tfrac{\|\mathbf{U}_{\|}^{(k)}\|}{\|\mathbf{U}\|} \geq \cos\theta^{(k)}, \tag{34}$$

where $\mathbf{S}^A(\mathbf{U})$ is the schedule chosen by the A-MW scheduler for given $\mathbf{U}$. Then, we obtain that, from $\langle \mathbf{1}, \mathbf{S}^A(\overline{\mathbf{U}})\rangle = 1$ and (16),

$$\begin{aligned}
&2E[\langle \mathbf{d}^{(k)}, \overline{\mathbf{U}}\rangle \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(\epsilon)} - \mathbf{1}\rangle] \\
&\quad = 2E[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\|}\| \cdot \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(\epsilon)}\rangle] \\
&\qquad - 2E[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\|}\| \cdot \langle \mathbf{d}^{(k)}, \mathbf{1}\rangle] \\
&\quad = 2E[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\|}\| \cdot \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle] \\
&\qquad + 2E[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\|}\| \cdot \epsilon \cdot \|\mathbf{d}^{(k)}\|] \\
&\qquad - 2E[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\|}\| \cdot b^{(k)}].
\end{aligned} \tag{35}$$

Combining the first term and the third term, we have

$$\begin{aligned}
&-2E\left[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\|}\| \cdot \Big(b^{(k)} - \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\Big)\right] \\
&= -2E\Big[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}\| \cdot \cos(\angle\overline{\mathbf{U}}, \overline{\mathbf{U}}_{\|}) \\
&\qquad \cdot \mathbb{1}_{\{\angle\overline{\mathbf{U}}, \overline{\mathbf{U}}_{\|} > \theta^{(k)}\}} \cdot \Big(b^{(k)} - \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\Big)\Big]
\end{aligned}$$

from $\|\overline{\mathbf{U}}_{\|}\| = \|\overline{\mathbf{U}}\| \cdot \cos(\angle\overline{\mathbf{U}}, \overline{\mathbf{U}}_{\|})$, and (34). Also, from $\|\overline{\mathbf{U}}_{\perp}\| = \|\overline{\mathbf{U}}\| \cdot \sin(\angle\overline{\mathbf{U}}, \overline{\mathbf{U}}_{\|})$, it can be drived further as

$$\begin{aligned}
&-2E\left[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\|}\| \cdot \Big(b^{(k)} - \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\Big)\right] \\
&\quad = -2E\Big[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\perp}\| \cdot \cot(\angle\overline{\mathbf{U}}, \overline{\mathbf{U}}_{\|}) \\
&\qquad \cdot \mathbb{1}_{\{\angle\overline{\mathbf{U}}, \overline{\mathbf{U}}_{\|} > \theta^{(k)}\}} \cdot \Big(b^{(k)} - \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\Big)\Big] \\
&\quad \geq -2E\Big[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\perp}\| \cdot \mathbb{1}_{\{\angle\overline{\mathbf{U}}, \overline{\mathbf{U}}_{\|} > \theta^{(k)}\}} \\
&\qquad \cdot \Big(b^{(k)} - \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\Big)\Big] \cdot \cot\theta^{(k)} \\
&\quad \geq -2E\Big[\|\mathbf{d}^{(k)}\| \cdot \|\overline{\mathbf{U}}_{\perp}\| \\
&\qquad \cdot \Big(b^{(k)} - \langle \mathbf{d}^{(k)}, \mathbf{S}^A(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\Big)\Big] \cdot \cot\theta^{(k)},
\end{aligned}$$

where the first inequality holds due to the decreasing property of cot in $(0, \pi/2]$. Note that from the concavity of the square

root, we have

$$
\begin{aligned}
&-2E\left[\|\overline{\mathbf{U}}_{\perp}\|\cdot\left(b^{(k)}-\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\right)\right]\\
&\geq -2\sqrt{E[\|\overline{\mathbf{U}}_{\perp}\|^{2}]\cdot E[(b^{(k)}-\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle)^{2}]}\\
&\geq -2\sqrt{\tfrac{\epsilon N_{2}^{(k)}}{\gamma^{(k)}}\cdot\|\boldsymbol{\lambda}^{(k)}\|\cdot\langle\mathbf{1},\boldsymbol{\lambda}^{(k)}\rangle\cdot\left((b^{(k)})^{2}+\langle\mathbf{d}^{(k)},\mathbf{1}/\boldsymbol{\lambda}^{(k)}\rangle^{2}\right)}
\end{aligned}
$$

where $\gamma^{(k)}:=\min\{b^{(k)}-\langle\mathbf{d}^{(k)},\mathbf{s}/\boldsymbol{\lambda}^{(k)}\rangle,\ \text{for all } \mathbf{s}\in\mathcal{S}\backslash\mathcal{F}^{(k)}\}$, the first inequality comes from the Hölder's inequatlity, and the last inequality holds due to Lemma 2 and the following lemma.

Lemma 4: Given $\boldsymbol{\lambda}^{(k)}$ and ϵ, if the arrival rate $\boldsymbol{\lambda}^{(\epsilon)}$ satisfies (16), then we have

$$1-\pi^{(k)}\leq\frac{\epsilon}{\gamma^{(k)}}\cdot\|\boldsymbol{\lambda}^{(k)}\|\cdot\langle\mathbf{1},\boldsymbol{\lambda}^{(k)}\rangle, \tag{36}$$

where $\pi^{(k)}:=P\{\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle=b^{(k)}\}$. It implies that

$$
\begin{aligned}
&E[\left(b^{(k)}-\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\right)^{2}]\\
&\quad=(1-\pi^{k})\cdot E\left[\left(b^{(k)}-\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\right)^{2}\right.\\
&\qquad\qquad\left.\mid\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\neq b^{(k)}\right]\\
&\quad\leq\frac{\epsilon}{\gamma^{(k)}}\cdot\|\boldsymbol{\lambda}^{(k)}\|\cdot\langle\mathbf{1},\boldsymbol{\lambda}^{(k)}\rangle\cdot\left((b^{(k)})^{2}+(\langle\mathbf{d}^{(k)},\mathbf{1}/\boldsymbol{\lambda}^{(k)}\rangle)^{2}\right).
\end{aligned}
$$

where the inequality holds from $S_{l}^{A}(\overline{\mathbf{U}})\leq 1$ for all l.

Proof: From the stability result of Lemma 1 and (22), we have

$$E[\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle]\geq b^{(k)}-\epsilon\|\boldsymbol{\lambda}^{(k)}\|\cdot\langle\mathbf{1},\boldsymbol{\lambda}^{(k)}\rangle,$$

which implies that

$$
\begin{aligned}
\pi^{(k)}\cdot b^{(k)}&+E[\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\cdot\mathbb{1}_{\{\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\neq b^{(k)}\}}]\\
&\geq b^{(k)}-\epsilon\|\boldsymbol{\lambda}^{(k)}\|\cdot\langle\mathbf{1},\boldsymbol{\lambda}^{(k)}\rangle.
\end{aligned}
$$

Also, note that

$$
\begin{aligned}
&E[\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\cdot\mathbb{1}_{\{\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\neq b^{(k)}\}}]\\
&\quad\leq(b^{(k)}-\gamma^{(k)})\cdot E[\mathbb{1}_{\{\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\neq b^{(k)}\}}]\\
&\quad=(b^{(k)}-\gamma^{(k)})\cdot(1-\pi^{(k)}).
\end{aligned}
$$

Combining the two inequalities, we obtain (36). ∎

Hence, from (35), we obtain that

$$
\begin{aligned}
&2E[\langle\mathbf{d}^{(k)},\overline{\mathbf{U}}\rangle\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(\epsilon)}-\mathbf{1}\rangle]\\
&\quad=2\epsilon\cdot\|\mathbf{d}^{(k)}\|^{2}\cdot E[\|\overline{\mathbf{U}}_{\|}\|]+O(\sqrt{\epsilon}).
\end{aligned}
\tag{37}
$$

We now consider a bound on the right side of (33):

$$
\begin{aligned}
&E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{X}\cdot\mathbf{S}^{A}(\overline{\mathbf{U}})\rangle^{2}]\\
&=E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle^{2}]\\
&\quad+2E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\\
&\qquad\cdot\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}-\mathbf{X}\cdot\mathbf{S}^{A}(\overline{\mathbf{U}})\rangle]\\
&\quad+E[\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}-\mathbf{X}\cdot\mathbf{S}^{A}(\overline{\mathbf{U}})\rangle^{2}].
\end{aligned}
\tag{38}
$$

For the first term, we have, from Lemma 4,

$$
\begin{aligned}
&E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle^{2}]\\
&\quad=E[\left(b^{(k)}-\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\right)^{2}]=O(\epsilon).
\end{aligned}
$$

For the second term, we have, from the independence of $\mathbf{X}$ and (16),

$$
\begin{aligned}
&2E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\cdot\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}-\mathbf{X}\cdot\mathbf{S}^{A}(\overline{\mathbf{U}})\rangle]\\
&\quad=2E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\\
&\qquad\cdot\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}-\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(\epsilon)}\rangle]\\
&\quad=-2E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}\rangle\cdot\langle\mathbf{1},\epsilon\cdot\mathbf{S}^{A}(\overline{\mathbf{U}})\cdot\|\mathbf{d}^{(k)}\|\rangle]\\
&\quad\leq 0.
\end{aligned}
$$

For the third term, we have

$$
\begin{aligned}
&E[\langle\mathbf{d}^{(k)},\mathbf{S}^{A}(\overline{\mathbf{U}})/\boldsymbol{\lambda}^{(k)}-\mathbf{X}\cdot\mathbf{S}^{A}(\overline{\mathbf{U}})\rangle^{2}]\\
&=E\left[\left(\textstyle\sum_{l}d_{l}^{(k)}\cdot S_{l}^{A}(\overline{\mathbf{U}})\cdot\left(\frac{1}{\lambda_{l}^{(k)}}-X_{l}\right)\right)^{2}\right]\\
&=E\left[\textstyle\sum_{l}\left(d_{l}^{(k)}\cdot S_{l}^{A}(\overline{\mathbf{U}})\cdot\left(\frac{1}{\lambda_{l}^{(k)}}-X_{l}\right)\right)^{2}\right]+O(\epsilon^{2})\\
&=E\left[\textstyle\sum_{l}\left(d_{l}^{(k)}\cdot S_{l}^{A}(\overline{\mathbf{U}})\cdot\left(\frac{1}{\lambda_{l}^{(\epsilon)}}-X_{l}\right)\right)^{2}\right]+O(\epsilon)+O(\epsilon^{2})\\
&=\textstyle\sum_{l}(d_{l}^{(k)})^{2}\cdot(S_{l}^{A}(\overline{\mathbf{U}}))^{2}\cdot(\sigma_{l}^{X})^{2}+O(\epsilon)+O(\epsilon^{2})\\
&\leq\langle(\mathbf{d}^{(k)})^{2},(\boldsymbol{\sigma}^{X})^{2}\rangle+O(\epsilon)+O(\epsilon^{2}),
\end{aligned}
$$

where the last inequality comes from $S_{l}^{A}(\overline{\mathbf{U}})\leq 1$ for all l. Then, (38) can be bounded as

$$E[\langle\mathbf{d}^{(k)},\mathbf{1}-\mathbf{X}\cdot\mathbf{S}^{A}(\overline{\mathbf{U}})\rangle^{2}]\leq\langle(\mathbf{d}^{(k)})^{2},(\boldsymbol{\sigma}^{X})^{2}\rangle+O(\epsilon). \tag{39}$$

From (33), (37), and (39), we have that

$$2\epsilon\cdot\|\mathbf{d}^{(k)}\|^{2}\cdot E[\|\overline{\mathbf{U}}_{\|}\|]+O(\sqrt{\epsilon})\leq\langle(\mathbf{d}^{(k)})^{2},(\boldsymbol{\sigma}^{X})^{2}\rangle+O(\epsilon).$$

Taking $\epsilon\to 0$, we can obtain that

$$\lim_{\epsilon\to 0}\epsilon E[\|\overline{\mathbf{U}}_{\|}\|]\leq\frac{1}{2}\cdot\langle(\frac{\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|})^{2},(\boldsymbol{\sigma}^{X})^{2}\rangle. \tag{40}$$

C. A Lower bound

Let L denote the number of queues in the original system. We consider a single-queue server, with bounded arrival $\alpha(t)$ and bounded service $\beta(t)$ with $\alpha=E[\alpha(1)]$, $\sigma_{\alpha}^{2}=var(\alpha(1))$, $\beta=E[\beta(1)]$, $\sigma_{\beta}^{2}=var(\beta(1))$. The queue length $\Phi(t)$ that evolves as

$$\Phi(t+1)=(\Phi(t)+\alpha(t)-\beta(t))^{+}.$$

Following the same line of analysis in [28], we consider the arrival $\alpha(t)$ and the service $\beta(t)$ with $\beta-\alpha=\epsilon_{0}$, and let $\Phi^{(\epsilon_{0})}(t)$ denote the associated queue length process. For any $\epsilon_{0}>0$, $\{\Phi^{(\epsilon_{0})}(t)\}$ is a positive Harris recurrent Markov Chain, and converges in distribution to a random variable $\overline{\Phi}^{(\epsilon_{0})}$ with

all bounded moments [30], satisfying

$$E[\Phi^{(\epsilon_0)}(t)] \geq \frac{\zeta^{(\epsilon)}}{2\epsilon_0} - \frac{B}{2}, \tag{41}$$

where $\zeta^{(\epsilon)} := \sigma_\alpha^2 + \sigma_\beta^2 + \epsilon_0$, and B is an upper bound on $\beta(t)$.

Setting $\alpha(t) = \langle \mathbf{d}^{(k)}, \mathbf{1} \rangle$ and $\beta^{(\epsilon_0)}(t) = \max_{\mathbf{S} \in \Lambda} \langle \mathbf{d}^{(k)}, \mathbf{X} \cdot \mathbf{S} \rangle$, we have $\Phi^{(\epsilon_0)}(t) \leq \langle \mathbf{d}^{(k)}, \mathbf{U}(t) \rangle$. Now let us limit our interest to the set of the schedulers that do not consider instantaneous interarrival times (i.e., *interarrival-time-agnostic* schedulers). In this case, $\mathbf{S}(t)$ will be independent of $\mathbf{X}(t)$, and it is sufficient to maximize $\langle \mathbf{d}^{(k)}, \mathbf{S}/\lambda^{(\epsilon)} \rangle = \langle \mathbf{d}^{(k)}, \mathbf{S}/\boldsymbol{\lambda}^{(k)} \rangle + \epsilon \cdot \|\mathbf{d}^{(k)}\| \cdot \langle \mathbf{d}^{(k)}, \mathbf{S}/\mathbf{d}^{(k)} \rangle = \langle \mathbf{d}^{(k)}, \mathbf{S}/\boldsymbol{\lambda}^{(k)} \rangle + \epsilon \cdot \|\mathbf{d}^{(k)}\|$. Hence, $\beta^{(\epsilon_0)}(t) = \boldsymbol{\lambda}^{(k)}$ is an optimal solution with $\epsilon_0 = \epsilon \cdot \|\mathbf{d}^{(k)}\|$.

Under the assumption that the variance $\sigma^2_{\beta^{(\epsilon_0)}}$ converges to σ_β^2 as $\epsilon_0 \to 0$, and from (41), we can obtain random variable $\overline{\Phi}^{(\epsilon_0)}$ that satisfies

$$\liminf_{\epsilon_0 \to 0} \epsilon_0 E[\overline{\Phi}^{(\epsilon_0)}] \geq \frac{\sigma_\beta^2}{2} = \frac{1}{2} \langle (\mathbf{d}^{(k)})^2, (\boldsymbol{\sigma}^{\mathbf{X}})^2 \cdot (\boldsymbol{\lambda}^{(k)})^2 \rangle. \tag{42}$$

It is a lower bound for the set of *interarrival-time-agonostic* schedulers.