Emulating Round-Robin for Serving Dynamic Flows over Wireless Fading Channels

Bin Li
Department of ECBE
University of Rhode Island
Kingston, Rhode Island, USA
binli@uri.edu

Atilla Eryilmaz
Department of ECE
The Ohio State University
Columbus, Ohio, USA
eryilmaz.2@osu.edu

R. Srikant
CSL and Department of ECE
UIUC
Urbana, Illinois, USA
rsrikant@illinois.edu

ABSTRACT

Motivated by the Internet of Things (IoT) and Cyber-Physical Systems (CPS), we consider dynamic wireless fading networks, where each incoming flow has a random service demand and leaves the system once its service request is completed. In such networks, one of the primary goals of network algorithm design is to achieve short-term fairness that characterizes how often each flow is served, in addition to the more traditional goals such as throughput-optimality and delay-insensitivity to the flow size distribution. In wireline networks, all of these desired properties can be achieved by the round-robin scheduling algorithm. In the context of wireless networks, a natural extension of round-robin scheduling has been developed in the last few years through the use of a counter called the Time-Since-Last-Service (TSLS) that keeps track of the time that passed since the last service time of each flow. However, the performance of this round-robin-like algorithm has been primarily studied in the context of persistent flows that continuously inject packets into the network and do not ever leave the network. The analysis of dynamic flow arrivals and departures is challenging since each individual flow experiences independent wireless fading and thus, flows cannot be served in a strict round-robin manner. In this paper, we overcome this difficulty by exploring the intricate dynamics of TSLS-based algorithm and show that flows are provided round-robin-like service with a very high probability. Consequently, we then show that our algorithm can achieve throughput-optimality. Moreover, through simulations, we demonstrate that the proposed TSLS-based algorithm also exhibits desired properties such as delay-insensitivity and excellent short-term fairness performance in the presence of dynamic flows over wireless fading channels.

CCS CONCEPTS

• **Networks → Network performance modeling**; **Network performance analysis**; • **Mathematics of computing → Markov processes**;

KEYWORDS

Wireless scheduling, flow-level dynamics, throughout-optimality, delay-insensitivity, short-term fairness.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
MobiHoc'20, July 2020, Shanghai, China
© 2020 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

ACM Reference Format:
Bin Li, Atilla Eryilmaz, and R. Srikant. 2020. Emulating Round-Robin for Serving Dynamic Flows over Wireless Fading Channels. In *Proceedings of ACM Mobile Ad Hoc Networking and Computing (MobiHoc'20)*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

In the last decade, there has been an increasing research interest in dynamic wireless fading networks, where each incoming flow has a random service demand and leaves the system once its service request is completed (commonly referred to as flow-level dynamic model). This is primarily motivated by IoT and CPS applications where many small devices are expected to generate varying chunks of intermittent data to be communicated over wireless fading channels to a common server. For example, in a smart-home IoT application, some wireless sensors are monitoring air conditions (e.g., temperature and humidity) while others are monitoring the safety of the house. Hence, each wireless sensor intermittently generates bursty data and requires the network algorithm to quickly serve the generated data. While the latter property can be achieved by periodically scheduling transmissions for each sensor, it is hard to adapt to unpredictable traffic patterns in the presence of wireless interference and channel fading. As such, in these applications, the goals of the network algorithm design are to (i) support as many flows as possible (i.e., maximize system throughput.); (ii) guarantee that the delay is insensitive to the flow size distribution and thus is robust to the burstiness of the network traffic; (iii) serve flows as regularly as possible (i.e., maximizing short-term fairness, measured by the mean and the standard deviation of the inter-service time of each flow, characterizing how often the flow is served.)

In wireline networks, the aforementioned goals can be achieved by round-robin and its variants, such as Weighted Fair Queueing (WFQ) [5]. A well known extension of WFQ to wireless networks was developed in [9], where service is provided for each flow based on how far ahead or behind it compared to ideal WFQ. However, the algorithm was primarily designed for ON-OFF channels and relies on hyperparameters that limit the amount by which a flow can get ahead or fall behind ideal WFQ. The closest prior work [8] on emulating round-robin is through the use of a counter called the Time-Since-Last-Service (TSLS), which keeps track of the time since the last service of each flow. However, [8] focused on wireless networks consisting of flows that persist in the network forever. Although the proposed algorithm in [8] can be slightly modified to adapt to the case of dynamic flows, its throughput has been notoriously difficult to study, let alone its delay and short-term fairness performance. With fading, each individual flow experiences independent channel fading and leaves the network once it receives the desired amount of service. Therefore, despite incorporating the TSLS counter into the decision, flows are not served in an exact round-robin manner due to the wireless channel fading. This greatly complicates our analysis.

In this paper, we are able to show that the TSLS-based scheduling algorithm continues to be throughput-optimal in the presence of dynamic flows and wireless channel fading. Traditional fluid-limit-based techniques (e.g., [4, 12]) for throughput analysis cannot capture the discontinuities in the dynamics of TSLS counters and thus, cannot be applied in our considered network setups. Instead, our proof explicitly explores the intricate dynamics of the proposed TSLS-based scheduling algorithm. In particular, we establish the following two facts: (i) If the maximum TSLS value or the TSLS value of a flow that receives the service is large enough, then flows arriving after that flow do not receive any service with a very high probability and thus, flows are served in a round-robin fashion with a very high probability; (ii) If the maximum age value is large enough, then our proposed TSLS-based algorithm performs similarly to the age-based policy, which has already been shown to be throughput-optimal in the presence of dynamic flows and wireless fading (see [13]). Here, the age of a flow is defined as the amount of the time the flow staying in the system since it joined. To the best of our knowledge, this is the first work to characterize the round-robin behavior in a probabilistic way and thus the proof may be of independent interest. Moreover, through simulation results, we demonstrate that the proposed TSLS-based algorithm also exhibits desired properties such as delay-insensitivity and excellent short-term fairness performance in the presence of dynamic flows over fading channels.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 introduces the system model. Section 4 describes the TSLS-based algorithm and shows its throughput-optimality. Section 5 presents simulation results to demonstrate that the proposed TSLS-based algorithm also exhibits delay-insensitivity and short-term fairness performance. Section 6 presents detailed proofs of the throughput-optimality of the proposed TSLS-based algorithm. Section 7 concludes this paper.

2 PRIOR WORK AND CONTEXT

To put our work in comparative perspective, we now provide an overview of prior on wireless scheduling broadly, and round-robin-like emulations for dynamic flows specifically, and further, provide a brief discussion of our algorithm design philosophy in the context of prior work.

a) Scheduling Design for Dynamic Flows: In the presence of dynamic flows, the well-known queue-length-based MaxWeight algorithm (e.g., [14, 15]) is not throughput-optimal (see [16]). The main reason is that the queue-length-based MaxWeight algorithm myopically serves a flow with the maximum product of the residual size of the dynamic flow and its corresponding channel rate, and thus the flows with small backlogs may stay in the network forever. In [16], the authors developed a Maximum-Channel-Rate-First (MCRF) policy that always serves a flow with the maximum channel rate and showed that the proposed MCRF policy is throughput-optimal. This is because the probability of at least one flow having the maximum channel rate is close to 1 when there are sufficiently many flows in the system, which implies that it does not waste any service and thus does not incur any throughput loss under the MCRF policy. However, the MCRF policy may yield poor short-term fairness performance. Indeed, consider an extreme case of two classes of dynamic flows, where each class of flows has the same deterministic channel rate. In such a case, the MCRF policy always first serves the high-channel-rate flows and then serves the low-channel-rate flows if there is no any high-channel-rate flows in the system, and thus it does not behave like round-robin, yielding poor short-term fairness. In another interesting work [13], the authors proposed an age-based policy that serves a flow with the maximum product of age of a flow and its corresponding channel rate, where the age of a flow is defined as the amount of time the flow staying in the system since it joined. While it achieves maximum system throughput, it suffers from both poor delay and short-term fairness performance. For example, in a non-fading scenario with uniform channel rate, the age-based policy serves flows in the First-Come-First-Serve (FCFS) manner, resulting in the sensitive delay performance and poor short-term fairness.

b) Wireless Round-Robin Emulations: In the earliest on round-robin-like algorithm design for wireless network [9], the authors proposed a variant of WFQ that heuristically limits the amount by which a flow would lead or lag behind a true WFQ scheduler. Another interesting line of work (see [1, 2]) generalized the ideas of processor-sharing in bandwidth sharing networks (e.g., [10, 11]) and developed balance fairness schedulers in wireless networks that exhibit delay-insensitivity property. More recently, in [8], a round-robin-like algorithm was proposed in wireless networks through the use of TSLS counter in the scheduling decision. However, all these works considered the case of persistent flows that continuously inject packets into the network and will never leave the system. Thus, they did not address round-robin-like algorithm design for dynamic flows over wireless fading channels, which is of practical interest to the growing IoT and CPS applications.

c) Our design philosophy: As mentioned above, the simplest way to maximize throughput in dynamic wireless networks with fading is to transmit a flow with the maximum channel rate. However, this policy can be grossly unfair to flows which rarely see good channel states. The other extreme is to always provide service to the flow with the maximum TSLS (recall, TSLS is time since last service), but this can result in poor throughput since the flow with the maximum TSLS can be in a poor channel state. So it is clear that one should tradeoff the benefits of the two approaches. A natural idea is to break ties among flows with the maximum rates in favor of those with maximum TSLS. But again this could be unfair for those flows which never see good channel states. So we adopt the policy in [8] which schedules users with the largest product of current rate and an appropriate function of TSLS. We note that the model in [8] is very different (cf. Section 1), and it is not obvious that the algorithm will perform well in the truly dynamic setup considered in this paper. The main contribution of this paper is to establish through a combination of analysis and simulation that the algorithm indeed performs well in the dynamic setting with wireless fading.

3 SYSTEM MODEL

We consider the operation of a base station that serves dynamically arriving flows with random workloads over independently wireless fading channels. In particular, we consider I different classes of flows, where each class of flows have different arrival and channel statistics. We assume that the system operates in *slotted time* with normalized slots $t \in \{0, 1, 2, ...\}$. Let $A_i[t]$ denote the number of class-i flows arriving in time slot t; where $A_i[t]$ is independently and identically distributed (i.i.d.) over time with mean λ_i, and $A_i[t] \leq A_i^{\max}$ for some positive finite number $A_i^{\max}$, $\forall t \geq 0$. We also use $\mathcal{A}_i[t]$ to denote the set of newly arriving class-i flows in time slot t. Also, we assume that newly arriving flows cannot be scheduled until the next time slot. We use $F_{i,j}[t]$ to denote the number of packets of newly arriving flow j of class-i in time slot t; where $F_{i,j}[t]$ is i.i.d. over flows with mean $\eta_i > 0$ and $F_{i,j}[t] \leq F_i^{\max}$ for some positive finite number $F_i^{\max}$, $\forall t \geq 0$. We assume that all flows have K different channel rates $c_1, c_2, \ldots, c_K$ with $0 = c_1 < c_2 < \ldots < c_K \triangleq c^{\max}$, where c_k is a positive integer denoting that at most c_k packets can be delivered in one time slot. Let $C_{i,j}[t]$ denote the channel rate of flow j of class-i in time slot t, and $C_{i,j}[t]$ is i.i.d. over time and independently distributed over flows with $\Pr\{C_{i,j}[t] = c_k\} = p_{i,k}$, $\forall i = 1, 2, \ldots, I$, $\forall k = 1, 2, \ldots, K$. We assume that each class of flows have a strictly positive probability of having a maximum channel rate, i.e., $p_{i,K} > 0, \forall i = 1, 2, \ldots, I$.

Let $\mathcal{N}_i[t]$ be the set of class-i flows in the system in time slot t. Due to the wireless interference, at most one flow can be served in each time slot. Let $S_{i,j}[t] = 1$ if flow j of class-i is scheduled in time slot t and $S_{i,j}[t] = 0$ otherwise. Let $\mathbf{S}[t] = (S_{i,j}[t], \forall j \in \mathcal{N}_i[t], i = 1, 2, \ldots, I)$ denote a feasible schedule, where at most one element is equal to one. We use $\mathcal{S}$ to denote the collection of all feasible schedules. Let $R_{i,j}[t]$ denote the number of *residual* packets of flow j of class-i in time slot t, which are awaiting service. The flow leaves the system once all its packets have been served, i.e., its residual flow size reduces to 0. Therefore, the dynamics of $R_{i,j}[t]$ for flow j of class-i can be written as:

$$R_{i,j}[t+1] = \max\left\{R_{i,j}[t] - S_{i,j}[t]\, C_{i,j}[t], 0\right\}. \qquad (1)$$

We use $\rho_i \triangleq \lambda_i \mathbb{E}\left[\lceil F_{i,j}[t]/c^{\max} \rceil\right]$ to denote the traffic intensity of class-i flows, where $\lceil x \rceil$ denotes the minimum integer no smaller than x. ρ_i measures the average minimum number of slots required to complete service requests of an incoming class-i flow. In this paper, we consider the policies under which the system evolves as a Markov Chain. We call the system *stable* if the underlying Markov Chain is positive recurrent. It has been shown in [16] that $\rho \triangleq \sum_{i=1}^{I} \rho_i \leq 1$ in order to keep the above system stable. We say that a scheduler is *throughput-optimal* if it achieves the system stability for any traffic intensity $\rho < 1$.

4 WIRELESS ROUND-ROBIN DESIGN

In this work, we are interested in developing provably efficient scheduling policies that can allocate the time-varying resources fairly amongst the dynamic flows so that: (i) good channel conditions can be opportunistically utilized without significantly delaying flows with bad channel statistics, leading to *throughput-optimality*; and (ii) flows with large sizes do not unfairly block flows with small sizes from completing, thereby yielding both *delay-insensitivity* to the flow size distribution and *short-term fairness*. All of these properties are critical for the service of dynamically generated service requests in a shared wireless system.

4.1 TSLS-Based Algorithm Design

We define a dynamic parameter that facilitates the description of our policy. Let $W_{i,j}[t]$ be a counter of flow j of class-i, called Time-Since-Last-Service (TSLS), to keep track of the time that passed since flow j of class-i was last served. In particular, $W_{i,j}[t]$ increases by 1 in each time slot when flow j of class-i does not get service, either because it is not scheduled or because it has zero channel rate, and drops to 0 otherwise. More precisely, the evolution of $W_{i,j}[t]$ is described as follows:

$$W_{i,j}[t+1] = \left(W_{i,j}[t] + 1\right)\left(1 - \mathbb{1}_{\{S_{i,j}[t] C_{i,j}[t] > 0\}}\right), \qquad (2)$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function.

To facilitate the flexibility in the algorithm design, we define a set of functions:

$$\mathcal{G} \triangleq \{f \in \mathcal{F} : \text{for any } k \geq 1, b \geq 0, \text{ there exists a constant } c > 0 \text{ such that } f(x) - c \leq f(kx + b) \leq f(x) + c, \forall x \geq 0\},$$

where $\mathcal{F}$ is the set of non-negative, non-decreasing, differentiable and concave functions $f(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{y \to \infty} f(y) = \infty$ and $f(0) = 0$. Some examples of functions that are in class $\mathcal{G}$ are $f(x) = \log(1 + x)$ and $f(x) = \log(1 + x)/g(x)$, where $g(x)$ is an arbitrary positive, non-decreasing, and differentiable function which makes $f(x)$ an non-decreasing and concave function. Then, the scheduler that we will study can be described as in Algorithm 1.

Algorithm 1 (TSLS-based Scheduling Algorithm). *In each time slot t, select a feasible schedule $\mathbf{S}^*[t] \in \mathcal{S}$ such that*

$$\mathbf{S}^*[t] \in \operatorname*{argmax}_{\mathbf{S}[t] \in \mathcal{S}} \sum_{i=1}^{I} \sum_{j \in \mathcal{N}_i[t]} f\left(W_{i,j}[t]\right) C_{i,j}[t] S_{i,j}[t], \qquad (3)$$

where $f \in \mathcal{G}$.

The TSLS-based scheduling algorithm is exactly the round-robin algorithm in the non-fading case with homogeneous channel rates. In the presence of heterogeneous wireless channel fading, our proposed TSLS-based Scheduler tends to serve a flow that possesses high TSLS value and high achievable channel rate. Since high TSLS value for a flow implies that it has not received service for a long time, prioritizing high TSLS yields round-robin-like behavior. Yet, the presence of the rate $C_{i,j}[t]$ in (3) also incorporates the channel conditions into the decision.

We note that a similar TSLS-based has also appeared in our earlier work [8] in the context of persistent flows that continuously inject packets into the network and never leave the system. In the presence of dynamic flows, each individual flow expriences an independent wireless fading and thus flows are not served in an exact round-robin manner under the TSLS-based scheduling algorithm. Therefore, the proof techniques in [8] cannot be applied in the case of dynamic flows and hence new proof strategies are required to establish the throughput-optimality of the TSLS-based

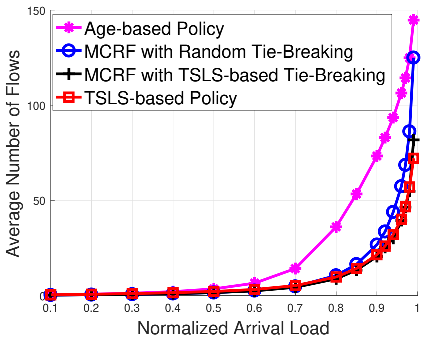


Figure 1: Throughput performance

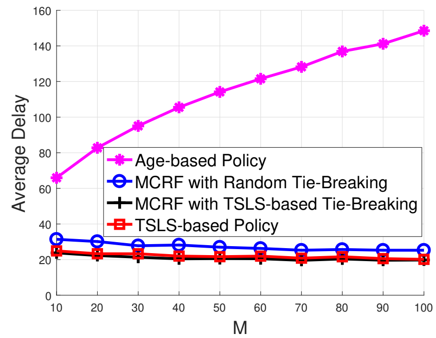


Figure 2: Average delay performance

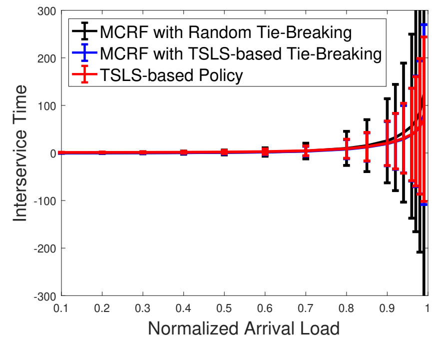


Figure 3: Short-term fairness

scheduling algorithm. In addition, the standard quadratic Lyapunov-based analyses as in [3] have unresolved technical flaws. Also, fluid-limit-based techniques (e.g., [4, 6, 13]) are difficult to capture the sharp dynamics of TSLS counters.

4.2 Main Result: Throughput-Optimality

In this section we present the main result that establishes the throughput-optimality of the TSLS-based scheduler presented in Algorithm 1. The most relevant work [8] to our work provides throughput-optimality of a TSLS-based design in the context of persistent flows. With the following result, we are now able to extend the setting of dynamic flows, where each flow independently experiences channel fading.

THEOREM 1. *The TSLS-based Scheduling Algorithm is throughput-optimal, i.e., it stabilizes the dynamic wireless fading network for any stabilizable traffic intensity* $\rho < 1$.

We note that this seemingly simple extension from persistent flows to dynamic flows presents significant new challenges to the analysis that demand substantially new arguments to establish. Accordingly, we dedicate Section 6 to present its proof. The proof explicitly explores the intricate dynamics of the TSLS-based scheduling algorithm and characterizes its round-robin behavior in a probabilistic way. Thus, the proof may be of independent interest in the analysis of other policies whereby fading and age-based schedulers are employed.

Subsequently, in Section 5, we will numerically demonstrate the delay-insensitivity and short-term fairness advantages of the TSLS-based algorithm due to its round-robin-like nature.

5 SIMULATIONS

In this section, we conduct simulations to compare our proposed TSLS-based scheduling policy with the logarithmic function (i.e., $f(x) = \log(1+x)$) to both age-based scheduling policy and Maximum-Channel-Rate-First (MCRF) policy with two different tie-breaking rules. Here, the age-based scheduling policy always serves a flow with the maximum product of age of the flow and its associated channel rate. The MCRF policy always serves a flow with the maximum channel rate with two tie-breaking rules: 1) random tie-breaking: if there are multiple flows achieving the maximum channel rate, then it randomly serves one of them; 2) TSLS-based tie-breaking: it serves the flow with the maximum TSLS value among all flows having the maximum channel rate.

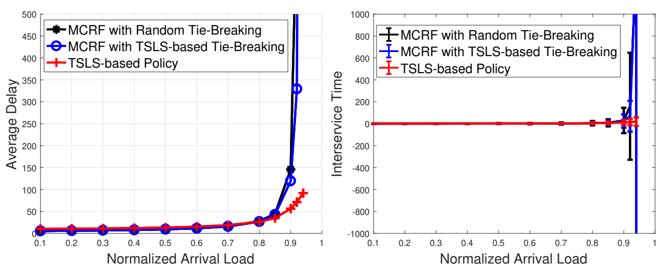
We consider two classes of flows, where each class of flows arrive at the system independently with Bernoulli distribution with the same rate λ. The flow size of the first class is equal to 3 with probability $1/2$ and 1 otherwise. The flow size of the second class has the following distribution: it is equal to M with probability $1/(M-1)$ and 1 otherwise, where $M \geq 2$ is some parameter that measures the burstiness of the flow size. Indeed, the mean flow size of the second class is always equal to 2 and its variance is equal to $(M-2)$, which linearly increases with the parameter M. Each flow of the first class has channel rates of 5 and 10 with corresponding probability of 0.01 and 0.99, respectively, while each flow of the second class has channel rates of 1 and 10 with corresponding probability of 0.99 and 0.01, respectively.

We evaluate the system performance in terms of throughput, mean delay, and short-term fairness. Fig. 1 shows the throughput performance when $M = 20$. In such a case, the throughput region is $\{\lambda : \lambda < 0.4872\}$ and thus we let $\lambda = 0.4872\theta$, where $\theta \in [0, 1)$ is called normalized arrival load. From Fig. 1, we can observe that all four policies stabilize the system for any $\theta \in [0, 1)$ in the above network setup, which validate the throughput-optimality of these four policies including our proposed TSLS-based policy.

In order to evaluate the delay-insensitivity performance, we vary M from 10 to 100 (i.e., the variance of the flow size of the second class linearly increases) when the normalized arrival load θ is 0.9. From Fig. 2, we can observe that the mean delay always keeps the same under both TSLS-based and MCRF policies with both tie-breaking rules, while it linearly increases under the age-based policy. This indicates that the delay performance under the age-based policy is sensitive to the variance of the flow size, which it is independent of variance of the flow size under both TSLS-based and MCRF policies. The reason lies in that the age-based policy roughly serves flows in a First-Come-First-Serve manner with a very high probability. In contrast, our proposed TSLS-based policy mimics the round-robin and thus yields delay-insensitive performance. Moreover, we can observe from Fig. 2 that both MCRF with TSLS-based tie-breaking and our TSLS-based policy outperforms the MCRF with random tie-breaking in terms of delay performance. This indicates the advantage of incorporating TSLS into the algorithm design.

Fig. 3 further compares the short-term fairness performance among TSLS-based policy and MCRF with two tie-breaking rules, where we plot the mean and 1.96 standard deviation (95% confidence interval) of the inter-service time of each flow. We can observe from Fig. 3 that our proposed TSLS-based policy outperforms the MCRF with random tie-breaking and performs slightly better than the MCRF with TSLS-based tie-breaking, especially in heavily loaded regimes.

In terms of heterogeneous maximum channel rates, our TSLS-based policy significantly outperforms MCRF with both tie-breaking rules. Indeed, consider two classes of flows with Bernoulli arrival processes. Each flow of the first class has the size of 5, 10, and 15 with corresponding probability 0.3, 0.5, and 0.2, while each flow of the second class has the same size with different probability distribution (0.6, 0.3, 0.1). Each flow of the first class has channel rates of 0, 1, 2, 5 and 10 with corresponding probability of 0.1, 0.2, 0.2, 0.2, and 0.3, respectively, while each flow of the second class has channel rates of 0 and 1 with corresponding probability of 0.2 and 0.8, respectively. The performance of different policies in this scenario is captured in the next figure.



(a) Average Delay Performance **(b) Short-term Fairness**

Figure 4: Performance comparison in the case of heterogeneous maximum channel rates

From Fig. 4a and Fig. 4b, we can observe that our TSLS-based policy significantly outperforms MCRF with both tie-breaking rules in terms of both delay and short-term fairness performance, especially in heavily loaded regimes. This is because that in terms of heterogeneous maximum channel rates, MCRF with both tie-breaking rules gave priority to flows with higher rates, while our TSLS-based still preserves round-robin nature and thus yields better delay and short-term fairness performance.

6 PROOF OF THROUGHPUT-OPTIMALITY

While throughput-optimality of the TSLS-based scheduler has been established in our earlier work [8] in the context of persistent flows, its analysis in presence of dynamic flows with wireless fading is highly non-trivial. Indeed, in the context of persistent flows, the throughout-optimality is built on the following two facts: (i) flows are served in an exact round-robin manner within each link under the TSLS-based scheduler; (ii) the policy performs similarly to the age-based policy across links. However, in the presence of dynamic flows, each individual flow experiences an independent channel fading and thus flows are not served in a round-robin manner anymore. In such a case, the arguments in [8] do not apply and new proof strategies are required to establish the throughout-optimality of the TSLS-based scheduler in the presence of dynamic flows.

In this section, we will prove Theorem 1. The proof is built on the following two facts: i) If the maximum TSLS value or the TSLS value of a flow that receives the service is large enough, then flows arriving after that flow do not receive any service with a very high probability and thus flows are served in a round-robin fashion with a very high probability (see Lemma 1); ii) If the maximum age value is large enough, then our TSLS-based algorithm performs similarly to the age-based policy (see Lemma 2), which has already been shown to be throughput-optimal in the case of homogeneous maximum channel rates (see [13]).

We first establish the first fact. Let $\widehat{j}[t]$ be the index of the flow that is served in time slot t or the flow with the maximum TSLS in time slot t. We use $\widehat{W}[t]$ to denote the TSLS value of flow $\widehat{j}[t]$, which implies that flow $\widehat{j}[t]$ was served in time slot $t-\widehat{W}[t]$. Then, we have the following lemma.

Lemma 1. *For any $\gamma \in (0,1)$, there exists a sufficiently large $J(\gamma) > 0$ such that given $\widehat{W}[t] = b \geq J(\gamma)$, with probability at least $1-\gamma$, that all flows arriving after time $t-(1-\gamma)\widehat{W}[t]-1$ are not served in the interval $[t-(1-\gamma)\widehat{W}[t], t-1]$, i.e.,*

$$\Pr\left\{ \bigcap_{\tau \in [t-(1-\gamma)\widehat{W}[t], t-1]} \mathcal{F}_\tau \middle| \widehat{W}[t] = b \right\} \geq 1-\gamma, \tag{4}$$

where $\mathcal{F}_\tau \triangleq \{$all flows arriving after time $t-(1-\gamma)\widehat{W}[t]-1$ are not served at time slot $\tau\}$, $\tau \in [t-(1-\gamma)\widehat{W}[t], t-1]$.

Proof. The proof consists of the following three steps: i) If $\widehat{W}[t]$ is large enough, with a very high probability, the amount of workload existing in time slot $t-\widehat{W}[t]$ is large enough; ii) Step i) ensures that flows arriving just slightly after $t-\widehat{W}[t]$ are still present in the system in time slot t; iii) Step ii) guarantees that flows arriving slightly after time $t-\widehat{W}[t]$ are not served before time slot t.

We will first show step i). In particular, we will show that for any $\gamma \in (0,1)$, there exists a sufficiently large $J_0(\gamma) > 0$ such that for any $b \geq J_0(\gamma)$, we have

$$\Pr\left\{ Q\left[t-\widehat{W}[t]\right] \geq \widehat{W}[t]-\kappa(b) \middle| \widehat{W}[t] = b \right\} \geq 1-\frac{\gamma}{2b^2}, \tag{5}$$

where $\kappa(b) \triangleq 1+\log_a \frac{2b^2}{\gamma\psi(a)}$, $\psi(a) \triangleq (1-(a-1)/p_{\widehat{i},K})^{(1-p_{\widehat{i},K})^2/p_{\widehat{i},K}}$, a is some constant between 1 and $1+p_{\widehat{i},K}$, and $\widehat{i}$ is the index of class that flow $\widehat{j}$ belongs to.

To that end, we first figure out the minimum amount of workload existing in the system in time slot $t-\widehat{W}[t]$ such that flow $\widehat{j}[t]$ cannot be served in the interval $[t-\widehat{W}[t], t]$. Assume that flow $\widehat{j}[t]$ has the maximum channel rate $c^{\max}$ for H_{ON} slots in the interval $[t-\widehat{W}[t]+1, t-1]$ with the length of $\widehat{W}[t]-1$. Among these H_{ON} slots, we call the one m^{th} closest to time slot t as the m^{th} "ON" slot, where $m = 1, 2, \ldots, H_{ON}$. All other slots between $t-\widehat{W}[t]+1$ and $t-1$ are called "OFF" slots. Fig. 6 illustrates "ON" and "OFF" slots of flow $\widehat{j}[t]$ in the interval $[t-\widehat{W}[t]+1, t-1]$.

In order to make sure that flow $\widehat{j}[t]$ is not served in the interval $[t-\widehat{W}[t]+1, t-1]$, the following two conditions should be satisfied:

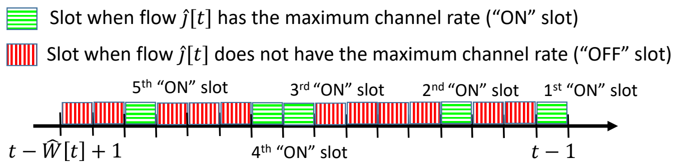


Figure 5: Illustration of channel states of flow $\widehat{j}[t]$ in the interval $[t-\widehat{W}[t]+1, t-1]$.

(i) There should be flows with at least m units of workload coming before time $t-\widehat{W}[t]+1$ that are present at the m^{th} "ON" slot.

(ii) At each "OFF" slot between the $(m+1)^{th}$ "ON" slot and the m^{th} "ON" slot, in order to meet the condition (i), there should be flows with at least one unit of workload coming before time $t-\widehat{W}[t]+1$ when at least one of flows with the total m units of workload has the maximum channel rate $c^{\max}$.

Otherwise, flow $\widehat{j}[t]$ would receive another service in one of "ON" slots in the interval $[t-\widehat{W}[t]+1, t-1]$ under the TSLS-based policy.

Let $H_{m,OFF}$ denote the number of slots between the $(m+1)^{th}$ "ON" slot and the m^{th} "ON" slot, where $m=1,\ldots,H_{ON}-1$. We use $H_{H_{ON},OFF}$ to denote the number of slots between time $t-\widehat{W}[t]+1$ and the H_{ON}^{th} "ON" slot. Therefore, in order to guarantee that flow $\widehat{j}[t]$ does not receive any service in the interval $[t-\widehat{W}[t]+1, t-1]$, the total amount of workload existing in the system in time slot $t-\widehat{W}[t]$, i.e., $Q\left[t-\widehat{W}[t]\right]$, should satisfy

$$\begin{aligned} Q\left[t-\widehat{W}[t]\right] &\geq H_{ON}+\sum_{m=1}^{H_{ON}}\sum_{\tau=1}^{H_{m,OFF}}\mathbb{1}_{\{\overline{\mathcal{H}}_m \text{ at slot } \tau\}} \\ &\geq H_{ON}+\sum_{m=1}^{H_{ON}}H_{m,OFF}-\sum_{m=1}^{H_{ON}}\sum_{\tau=1}^{H_{m,OFF}}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}} \\ &=\widehat{W}[t]-1-\sum_{m=1}^{H_{ON}}\sum_{\tau=1}^{H_{m,OFF}}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}, \end{aligned} \tag{6}$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function, $\mathcal{H}_m$ denotes the event that all flows with the total m units of workload do not have the maximum channel rate $c^{\max}$, $\overline{\mathcal{H}}_m$ is the complement of event $\mathcal{H}_m$, and the last step is true since the interval $[t-\widehat{W}[t]+1, t-1]$ with the length of $\widehat{W}[t]-1$ contains of H_{ON} "ON" slots and $\sum_{m=1}^{H_{ON}}H_{m,OFF}$ "OFF" slots.

Next, we will show that for any $\gamma\in(0,1)$, we have

$$\Pr\left\{\sum_{m=1}^{H_{\text{ON}}}\sum_{\tau=1}^{H_{m,\text{OFF}}}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}\geq\kappa(b)-1\middle|\widehat{W}[t]=b\right\}\leq\frac{\gamma}{2b^2}, \tag{7}$$

where we recall that $\kappa(b)\triangleq 1+\log_a\frac{2b^2}{\gamma\psi(a)}$, $\psi(a)\triangleq(1-(a-1)/p_{\widehat{i},K})^{(1-p_{\widehat{i},K})^2/p_{\widehat{i},K}}$, and a is some constant between 1 and $1+p_{\widehat{i},K}$. This and (6) together imply (5).

Next, we show inequality (7) to complete the proof for the step i). Noting that $H_{m,\text{OFF}}, m=1,2,\cdots,H_{\text{ON}}-1$ are i.i.d. geometrically distributed with parameter $p_{\widehat{i},K}$, we have

$$\begin{aligned} &\Pr\left\{\sum_{m=1}^{H_{\text{ON}}}\sum_{\tau=1}^{H_{m,\text{OFF}}}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}\geq\kappa(b)-1\middle|\widehat{W}[t]=b\right\} \\ &\overset{(a)}{\leq}\Pr\left\{\sum_{m=1}^{\widehat{W}[t]-1}\sum_{\tau=1}^{Z_m}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}\geq\kappa(b)-1\middle|\widehat{W}[t]=b\right\} \\ &\overset{(b)}{=}\Pr\left\{a^{\sum_{m=1}^{\widehat{W}[t]-1}\sum_{\tau=1}^{Z_m}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}}\geq a^{\kappa(b)-1}\middle|\widehat{W}[t]=b\right\} \\ &\overset{(c)}{\leq}a^{-(\kappa(b)-1)}\mathbb{E}\left[a^{\sum_{m=1}^{\widehat{W}[t]-1}\sum_{\tau=1}^{Z_m}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}}\middle|\widehat{W}[t]=b\right], \end{aligned} \tag{8}$$

where step (a) is true for that $Z_m\in\{0,1,2,\cdots\}, m=1,2,\cdots$, are i.i.d. geometrically distributed with parameter $p_{\widehat{i},K}$; (b) is true for some constant $a\in(1,1+p_{\widehat{i},K})$; (c) uses Markov's Inequality.

Next, we consider the second component in (8).

$$\begin{aligned} &\mathbb{E}\left[a^{\sum_{m=1}^{\widehat{W}[t]-1}\sum_{\tau=1}^{Z_m}\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}}\middle|\widehat{W}[t]=b\right] \\ &\overset{(a)}{=}\prod_{m=1}^{b-1}\mathbb{E}\left[\prod_{\tau=1}^{Z_m}a^{\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}}\right] \\ &\overset{(b)}{=}\prod_{m=1}^{b-1}\sum_{l=0}^{\infty}p_{\widehat{i},K}(1-p_{\widehat{i},K})^i\prod_{\tau=1}^{l}\mathbb{E}\left[a^{\mathbb{1}_{\{\mathcal{H}_m \text{ at slot } \tau\}}}\right] \\ &\overset{(c)}{=}\prod_{m=1}^{b-1}\sum_{l=0}^{\infty}p_{\widehat{i},K}\prod_{\tau=1}^{l}\left(a(1-p_{\widehat{i},K})^m+1-(1-p_{\widehat{i},K})^m\right) \\ &=\prod_{m=1}^{b-1}p_{\widehat{i},K}\sum_{l=0}^{\infty}\left((1-p_{\widehat{i},K})\left((a-1)(1-p_{\widehat{i},K})^m+1\right)\right)^l \\ &\overset{(d)}{=}\prod_{m=1}^{b-1}\frac{p_{\widehat{i},K}}{1-(1-p_{\widehat{i},K})\left((a-1)(1-p_{\widehat{i},K})^m+1\right)} \\ &=\frac{1}{\prod_{m=1}^{b-1}\left(1-\frac{a-1}{p_{\widehat{i},K}}(1-p_{\widehat{i},K})^{m+1}\right)}, \end{aligned} \tag{9}$$

where step (a) uses the fact that $Z_m, m=1,2,\ldots,$ are i.i.d.; (b) is true since $Z_m\in\{0,1,2,\ldots\}$ is geometrically distributed with parameter $p_{\widehat{i},K}$; (c) follows from the definition of event $\mathcal{H}_m$; (d) utilizes the fact that

$$0<(1-p_{\widehat{i},K})\left((a-1)(1-p_{\widehat{i},K})+1\right)<1,$$

since $a<1+p_{\widehat{i},K}$.

By using the inequality $1-ux\geq(1-u)^x$ for any $0<u<1$ and $0\leq x\leq 1$, and noting the fact that $(a-1)/p_{\widehat{i},K}<1$, we have

$$\begin{aligned} \prod_{m=1}^{b-1}\left(1-\frac{a-1}{p_{\widehat{i},K}}(1-p_{\widehat{i},K})^{m+1}\right)&\geq\left(1-\frac{a-1}{p_{\widehat{i},K}}\right)^{\sum_{m=1}^{b-1}(1-p_{\widehat{i},K})^{m+1}} \\ &\overset{(a)}{\geq}\left(1-\frac{a-1}{p_{\widehat{i},K}}\right)^{\sum_{m=1}^{\infty}(1-p_{\widehat{i},K})^{m+1}} \\ &\overset{(b)}{=}\psi(a), \end{aligned} \tag{10}$$

where step (a) uses the fact that $(1-u)^x$ is non-increasing with respect to x for any $0 < u < 1$; (b) is true for $\psi(a) = (1-(a-1)/p_{\widehat{i},K})^{(1-p_{\widehat{i},K})^2/p_{\widehat{i},K}}$.

By combining (8), (9) and (10), we have

$$\Pr\left\{\sum_{m=1}^{H_{\text{ON}}}\sum_{k=1}^{H_{m,\text{OFF}}} 1_{\{\mathcal{H}_m \text{ at slot } k\}} \geq \kappa(b)-1 \middle| \widehat{W}[t]=b\right\} \leq \frac{1}{\psi(a)}a^{-(\kappa(b)-1)} = \frac{\gamma}{2b^2}, \tag{11}$$

where the last step follows from the definition of $\kappa(b) = 1 + \log_a \frac{2b^2}{\gamma\psi(a)}$.

Next, we will show step ii), i.e., if $\widehat{W}[t]$ is large enough, then with a very high probability, flows arriving just slightly after flow $\widehat{j}[t]$ is still present in the system in time slot t. In particular, for any $\gamma \in (0,1)$, there exists a sufficiently large $J(\gamma) > 0$ such that for any $b \geq J(\gamma)$,

$$\Pr\left\{Q(t-(1-\gamma)\widehat{W}[t],t) \geq Q_0(\widehat{W}[t]) | \widehat{W}[t]=b\right\} \geq 1-\frac{\gamma}{b^2}, \tag{12}$$

where $Q_0(x) \triangleq \nu^{\max}\ln(\gamma/x^2)/\ln(1-p_{\min})$ or equivalently $(1-p_{\min})^{Q_0(x)/\nu^{\max}} = \gamma/x^2$, for any $x \geq 1$, $\nu^{\max} \triangleq \max_i\lceil F_i^{\max}/c^{\max}\rceil$, and $p_{\min} \triangleq \min_{i=1,\ldots,I} p_{i,K}$.

Since at most one unit of workload can be served in each time slot, at most $\widehat{W}[t]$ amount of workload can be reduced between $t-\widehat{W}[t]$ and $t-1$. Thus, the total amount of workload arriving before time $t-(1-\gamma)\widehat{W}[t]$ that are still present in the system in time slot t should have the following relationship:

$$\begin{aligned} &Q\left(t-(1-\gamma)\widehat{W}[t],t\right) \\ \geq &Q[t-\widehat{W}[t]] + \sum_{\tau=t-\widehat{W}[t]+1}^{t-(1-\gamma)\widehat{W}[t]}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil - \widehat{W}[t] \\ \geq &\sum_{\tau=t-\widehat{W}[t]+1}^{t-(1-\gamma)\widehat{W}[t]}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil - \kappa(b), \end{aligned} \tag{13}$$

holds with a probability at least $1-\gamma/(2b^2)$, where the last step directly follows from step i) (cf. (5)). This implies that

$$\begin{aligned} &\Pr\left\{Q(t-(1-\gamma)\widehat{W}[t],t) \geq Q_0(\widehat{W}[t]) \middle| \widehat{W}[t]=b\right\} \\ \geq &\Pr\left\{\sum_{\tau=t-\widehat{W}[t]+1}^{t-(1-\gamma)\widehat{W}[t]}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil \geq Q_0(\widehat{W}[t])+\kappa(b)\middle|\widehat{W}[t]\right\} \\ &\cdot\left(1-\frac{\gamma}{2b^2}\right) \\ = &\Pr\left\{\sum_{\tau=1}^{\gamma\widehat{W}[t]}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil \geq Q_0(\widehat{W}[t])+\kappa(b)\middle|\widehat{W}[t]\right\} \\ &\cdot\left(1-\frac{\gamma}{2b^2}\right) \\ = &\Pr\left\{\sum_{\tau=1}^{\gamma b}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil \geq Q_0(b)+\kappa(b)\right\}\left(1-\frac{\gamma}{2b^2}\right), \end{aligned} \tag{14}$$

where the second last step is true since $\mathcal{A}_i[t]$ and $F_{i,j}[t]$, $t \geq 0$, are i.i.d.. In addition, according to the Hoeffding's inequality, we have

$$\begin{aligned} &\Pr\left\{\sum_{\tau=1}^{\gamma b}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil < \frac{1}{2}\rho\gamma b\right\} \\ = &\Pr\left\{\sum_{\tau=1}^{\gamma b}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil - \rho\gamma b < -\frac{1}{2}\rho\gamma b\right\} \\ \leq &\exp\left(-\frac{\rho^2\gamma b}{2(KA^{\max}\nu^{\max})^2}\right) \leq \frac{\gamma}{2b^2}, \end{aligned} \tag{15}$$

where the last inequality holds for $b \geq J_1(\gamma)$ and $J_1(\gamma) > 0$ is some sufficiently large number depending on γ. This implies that

$$\Pr\left\{\sum_{\tau=1}^{\gamma b}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil \geq \frac{1}{2}\lambda\gamma b\right\} \geq 1-\frac{\gamma}{2b^2}. \tag{16}$$

In addition, there exists a $J_2(\gamma)$ such that for any $b > J_2(\gamma)$, we have

$$\frac{1}{2}\lambda\gamma b \geq Q_0(b)+\kappa(b), \tag{17}$$

where we use the fact that $Q_0(b)$ is a logarithmic function of b, and the definition of $\kappa(b)$.

By combining (14), (16), and (17), we have that for any $b \geq J(\gamma) \triangleq \max\{J_1(\gamma), J_2(\gamma)\}$,

$$\begin{aligned} &\Pr\left\{Q(t-(1-\gamma)\widehat{W}[t],t) \geq Q_0(\widehat{W}[t]) \middle| \widehat{W}[t]=b\right\} \\ \overset{(a)}{\geq} &\Pr\left\{\sum_{\tau=1}^{\gamma b}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil \geq Q_0(b)+\kappa(b)\right\}\left(1-\frac{\gamma}{2b^2}\right) \\ \overset{(b)}{\geq} &\Pr\left\{\sum_{\tau=1}^{\gamma b}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i[\tau]}\left\lceil\frac{F_{i,j}[\tau]}{c^{\max}}\right\rceil \geq \frac{1}{2}\lambda\gamma b\right\}\left(1-\frac{\gamma}{2b^2}\right) \\ \overset{(c)}{\geq} &\left(1-\frac{\gamma}{2b^2}\right)^2 \geq 1-\frac{\gamma}{b^2}, \end{aligned} \tag{18}$$

where the step (a) follows from (14); (b) uses (17); (c) follows from (16).

Building on both steps i) and ii), we are ready to prove step iii) (cf., (4)). For some $\tau \in [t-(1-\gamma)\widehat{W}[t], t-1]$, we have

$$\begin{aligned} &\Pr\left\{\overline{\mathcal{F}_\tau}\middle|\widehat{W}[t]=b\right\} \\ \overset{(a)}{\leq} &\Pr\{\text{all flows in } \mathcal{N}(t-(1-\gamma)\widehat{W}[t],\tau) \text{ do not have the rate} \\ &\quad c^{\max}|\widehat{W}[t]=b\} \\ \overset{(b)}{\leq} &\mathbb{E}\left[(1-p_{\min})^{Q(t-(1-\gamma)\widehat{W}[t],\tau)/\nu^{\max}}\middle|\widehat{W}[t]=b\right] \\ \overset{(c)}{\leq} &\mathbb{E}\left[(1-p_{\min})^{Q(t-(1-\gamma)\widehat{W}[t],t)/\nu^{\max}}\middle|\widehat{W}[t]=b\right] \end{aligned}$$

$$\overset{(d)}{=}\mathbb{E}\left[(1-p_{\min})^{Q(t-(1-\gamma)\widehat{W}[t],t)/v^{\max}}1_{\mathcal{E}}\middle|\widehat{W}[t]=b\right]$$
$$+\mathbb{E}\left[(1-p_{\min})^{Q(t-(1-\gamma)\widehat{W}[t],t)/v^{\max}}1_{\overline{\mathcal{E}}}\middle|\widehat{W}[t]=b\right]$$
$$\overset{(e)}{\leq}\mathbb{E}\left[\frac{\gamma}{(\widehat{W}[t])^2}1_{\mathcal{E}}\middle|\widehat{W}[t]=b\right]+\Pr\left\{\overline{\mathcal{E}}\middle|\widehat{W}[t]=b\right\}$$
$$\overset{(f)}{\leq}\frac{2\gamma}{b^2}, \tag{19}$$

where step (a) follows from the fact that if at least one flow in $\mathcal{N}(t-(1-\gamma)\widehat{W}[t],\tau)$ has the maximum channel rate $c_{\max}$, then the flow coming after time $t-(1-\gamma)\widehat{W}[t]$ cannot be served in time slot τ under the TSLS-based policy (i.e., event $\mathcal{F}_\tau$ happens); (b) uses the fact that the number of flows in $\mathcal{N}(t-(1-\gamma)\widehat{W}[t],\tau)$ is at least $Q(t-(1-\gamma)\widehat{W}[t],\tau)/v^{\max}$, $v^{\max}=\max_i\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil$, and $p_{\min}\triangleq\min_{i=1,\ldots,I}p_{i,K}$; (c) uses the fact that $Q(\tau,t)$ is non-increasing in t by its definition and the fact that $(1-p_{\min})^x$ is non-increasing in x; (d) is true for the event $\mathcal{E}\triangleq\{Q(t-(1-\gamma)\widehat{W}[t],t)<Q_0(\widehat{W}[t])\}$; (e) follows from the fact that $(1-p_{\min})^x$ is non-increasing and the definition of $Q_0(x)$; (f) uses inequality (12).

Therefore, we have

$$\Pr\left\{\bigcup_{\tau\in[t-(1-\gamma)\widehat{W}[t],t-1]}\overline{\mathcal{F}_\tau}\middle|\widehat{W}[t]=b\right\}$$
$$\overset{(a)}{\leq}\sum_{\tau\in[t-(1-\gamma)b,t-1]}\Pr\left\{\overline{\mathcal{F}_\tau}\middle|\widehat{W}[t]=b\right\}$$
$$\overset{(b)}{\leq}\frac{2\gamma}{b}\overset{(c)}{\leq}\gamma, \tag{20}$$

where step (a) uses the union bound; (b) uses (19); (c) holds since $J(\gamma)$ is sufficiently large and $b\geq J(\gamma)$. Hence, we have the desired result. □

In order to establish the second fact that the TSLS-based policy performs similarly to the age-based policy when the maximum age of he flow is large enough. We introduce $T_{i,j}[t]$ to denote the age of flow j of class-i in time slot t, whereby $T_{i,j}[t]$ starts from 0 at its arrival and is incremented (by 1) in each time slot until flow j of class-i leaves the system. More precisely, the evolution of $T_{i,j}[t]$ can be written as

$$T_{i,j}[t+1]=\left(T_{i,j}[t]+1\right)\left(1-\mathbb{1}_{\{R_{i,j}[t+1]=0\}}\right), \tag{21}$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function.

Lemma 2. *For any $\gamma\in(0,1)$, we have*

$$\Pr\left\{f(T^{\max}[t])\leq f(W^{\max}[t])+G(\gamma)\middle|T^{\max}[t]\right\}\geq(1-\gamma)^2,$$

where $G(\gamma)$ is some positive value depending on γ, $f\in\mathcal{G}$, $\zeta=1/\left((1+1/(\lambda(1-\gamma)))^{F^{\max}}-1\right)$, $F^{\max}=\max_i F_i^{\max}$, and $\lambda\triangleq\sum_{i=1}^{I}\lambda_i$.

Proof. Note that flows at time t consist of flows arriving between $t-W^{\max}[t]$ and $t-1$ and flows that have already existed in time slot $t-W_{\max}[t]$. Since there are at most $A^{\max}$ (recall that $A^{\max}=\max_i A_i^{\max}$) flows arriving at the system in each time slot, the number of flows arriving between $t-W^{\max}[t]$ and t is at most $A^{\max}W^{\max}[t]$. In addition, there are at most $W^{\max}[t]+1$ flows that have already existed in time slot $t-W^{\max}[t]$, because i) at most one existing flow has TSLS value of zero since it received service in time slot $t-W^{\max}[t]-1$; (ii) other existing flows must get service at least once between $t-W^{\max}[t]$ and $t-1$. Otherwise, at least one of these flows has TSLS value greater than $W^{\max}[t]$. This implies that

$$|\mathcal{N}[t]|\leq(A^{\max}+1)W^{\max}[t]+1,\forall t\geq 0, \tag{22}$$

holds for any sample path, where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$.

Next, we will show that if $T^{\max}[t]\geq\widehat{J}(\gamma)/\zeta$, then we have

$$\Pr\left\{|\mathcal{N}[t]|\geq\frac{1}{2}\lambda(1-\gamma)\zeta T^{\max}[t]\middle|T^{\max}[t]\right\}\geq(1-\gamma)^2. \tag{23}$$

We use $j^{\max}[t]$ to denote the index of one of flows with the maximum age of $T^{\max}[t]$ in time slot t. Since each flow has at most $F^{\max}$ (recall that $F^{\max}=\max_i F_i^{\max}$) packets, flow $j^{\max}[t]$ cannot be served more than $F^{\max}-1$ times between $t-T^{\max}[t]$ and t in order for it to stay in the system in time slot t. Without loss of generality, we assume that flow $j^{\max}[t]$ was served K times in the interval $[t-T^{\max}[t],t]$, where $K=0,1,2,\ldots,F^{\max}-1$. In particular, we assume that flow $j^{\max}[t]$ was served in time slots $t_1,t_2,\ldots,t_K$, where $t_1>t_2>\ldots>t_K$. Let $t_0=t-1$ and $t_{K+1}=t-T^{\max}[t]$. Therefore, the interval $[t-T^{\max}[t],t-1]$ is partitioned into $K+1$ subintervals (see Fig. 6); the k^{th} subinterval $[t_k,t_{k-1}]$ has length of $a_kT^{\max}[t]$ ($\forall k=1,2,\ldots,K+1$), where $a_1,a_2,\ldots,a_K,a_{K+1}\geq 0$ and $\sum_{k=1}^{K+1}a_k=1$.

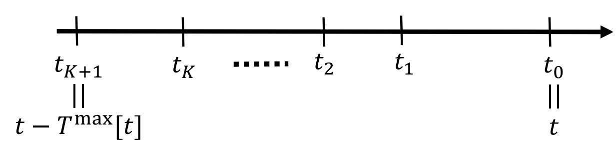


Figure 6: Time interval partition: flow $j^{\max}[t]$ was served in time slots $t_1,t_2,\ldots,t_K$

Hence, if $\zeta=1/\left((1+1/\beta)^{F^{\max}}-1\right)$ for some $\beta>0$, then, at least one of the following inequalities should hold:

$$\beta a_1\geq\zeta,$$
$$\beta a_2-a_1\geq\zeta,$$
$$\ldots,$$
$$\beta a_{K+1}-\sum_{k=1}^{K}a_k\geq\zeta. \tag{24}$$

Indeed, if all inequalities do not hold, then we have

$$\beta a_m-\sum_{k=1}^{m-1}a_k<\zeta,\forall m=1,2,\ldots,K+1. \tag{25}$$

By using mathematical induction, it is easy to show

$$\beta a_m<\zeta\left(1+\frac{1}{\beta}\right)^{m-1},\forall m=1,2,\ldots,K+1. \tag{26}$$

By summing the above inequalities over $m = 1, 2, \ldots, K+1$ and utilizing the fact that $\sum_{k=1}^{K+1} a_k = 1$, we have

$$\beta < \zeta \sum_{m=1}^{K+1} \left(1+\frac{1}{\beta}\right)^{m-1} = \zeta \frac{1-\left(1+\frac{1}{\beta}\right)^{K}}{1-\left(1+\frac{1}{\beta}\right)}. \tag{27}$$

Hence, we have

$$\frac{1}{\left(1+\frac{1}{\beta}\right)^{K}-1} < \zeta, \tag{28}$$

which contradicts with the fact that $\zeta = 1/\left(\left(1+\frac{1}{\beta}\right)^{F_{\max}}-1\right)$, because $1/\left((1+1/\beta)^K - 1\right)$ strictly decreases with respect to K and $K = 0, 1, \ldots, F_{\max}-1$.

Let k^* be the smallest index such that

$$\beta a_{k^*} - \sum_{k=1}^{k^*-1} a_k \geq \zeta, \tag{29}$$

where we set $\beta = \lambda(1-\gamma)/2$ and thus we have

$$\zeta = 1/\left((1+1/(\lambda(1-\gamma)/2))^{F^{\max}}-1\right).$$

Consider the interval $[t_{k^*}, t_{k^*-1}]$. Its length is $a_{k^*}T^{\max}[t]$. In such a case, given any $\gamma \in (0,1)$, according to Lemma 1, if the length $a_{k^*}T^{\max}[t]$ is large enough, with a very high probability, flows arriving between $t_{k^*}+(1-\gamma)a_{k^*}T^{\max}[t]$ and t_{k^*-1} cannot be served. This implies that with a very high probability, the number of flows in time slot t is at least $\sum_{i=1}^{I}\sum_{\tau=t_{k^*}+(1-\gamma)a_{k^*}T^{\max}[t]}^{t_{k^*-1}} A_i[\tau] - \sum_{k=1}^{k^*-1} a_k T^{\max}[t]$ due to the fact that at most one flow can leave the system in each time slot, i.e.,

$$\Pr\left\{|\mathcal{N}[t]| \geq \sum_{i=1}^{I}\sum_{\tau=t_{k^*}+(1-\gamma)a_{k^*}T^{\max}[t]}^{t_{k^*-1}} A_i[\tau] - \sum_{k=1}^{k^*-1} a_k T^{\max}[t] \middle| a_{k^*}T^{\max}[t] \geq J(\gamma)\right\} \geq 1-\gamma. \tag{30}$$

Given $a_{k^*}T^{\max}[t] = d$, according to the Hoeffding's inequality, we have

$$\begin{aligned}
&\Pr\left\{\sum_{i=1}^{I}\sum_{\tau=t_{k^*}+(1-\gamma)d}^{t_{k^*-1}} A_i[\tau] \leq \frac{\lambda}{2}(1-\gamma)d\right\} \\
=&\Pr\left\{\sum_{i=1}^{I}\sum_{\tau=1}^{(1-\gamma)d} A_i[\tau] \leq \frac{\lambda}{2}(1-\gamma)d\right\} \\
=&\Pr\left\{\sum_{i=1}^{I}\sum_{\tau=1}^{(1-\gamma)d} A_i[\tau] - \lambda(1-\gamma)d \leq -\frac{\lambda}{2}(1-\gamma)d\right\} \\
\leq&\exp\left(-\frac{\lambda^2(1-\gamma)d}{2(IA^{\max})^2}\right) \\
\leq&\gamma,
\end{aligned} \tag{31}$$

where the last inequality holds for $a_{k^*}T^{\max}[t] = d \geq J_3(\gamma)$ and $J_3(\gamma) > 0$ depends on γ.

In the rest of the proof, we omit the time index associated with $T^{\max}[t]$ and $W^{\max}[t]$ due to space restrictions. Hence, we have

$$\Pr\left\{|\mathcal{N}[t]| \geq \frac{1}{2}\lambda(1-\gamma)a_{k^*}T^{\max} - \sum_{k=1}^{k^*-1} a_k T^{\max} \middle| a_{k^*}T^{\max} \geq \widehat{J}(\gamma)\right\} \geq (1-\gamma)^2, \tag{32}$$

where $\widehat{J}(\gamma) \triangleq \max\{J(\gamma), J_3(\gamma)\}$.

Hence, according to the definition of k^* (see (29)), we have

$$\Pr\left\{|\mathcal{N}[t]| \geq \zeta T^{\max} \middle| T^{\max} \geq \widehat{J}(\gamma)/\zeta\right\} \geq (1-\gamma)^2, \tag{33}$$

where we use the fact that $a_{k^*} \geq \zeta$ (since $\beta < 1$).

By using inequality (22), we have

$$\begin{aligned}
&\Pr\left\{(A^{\max}+1)W^{\max}+1 \geq \zeta T^{\max} \middle| T^{\max} \geq \widehat{J}(\gamma)/\zeta\right\} \\
&\geq (1-\gamma)^2,
\end{aligned} \tag{34}$$

which implies

$$\begin{aligned}
&\Pr\left\{\zeta T^{\max}\mathbb{1}_{\{T^{\max}\geq \widehat{J}(\gamma)/\zeta\}} \leq (A^{\max}+1)W^{\max}+1 \middle| T^{\max}\right\} \\
&\geq (1-\gamma)^2,
\end{aligned} \tag{35}$$

If $\zeta T^{\max}\mathbb{1}_{\{T^{\max}\geq \widehat{J}(\gamma)/\zeta\}} \leq (A^{\max}+1)W^{\max}+1$, we have

$$T^{\max}\mathbb{1}_{\{T^{\max}\geq \widehat{J}(\gamma)/\zeta\}} \leq \frac{A^{\max}+1}{\zeta}W^{\max} + \frac{1}{\zeta}, \tag{36}$$

which implies that

$$\begin{aligned}
T^{\max} =& T^{\max}\mathbb{1}_{\{T^{\max}\geq \widehat{J}(\gamma)/\zeta\}} + T^{\max}\mathbb{1}_{\{T^{\max}< \widehat{J}(\gamma)/\zeta\}} \\
\leq& \frac{A^{\max}+1}{\zeta}W^{\max} + \frac{1}{\zeta} + \frac{\widehat{J}(\gamma)}{\zeta}.
\end{aligned} \tag{37}$$

Since $f \in \mathcal{G}$, there exists a $G(\gamma)$ such that

$$f(T^{\max}) \leq f(W^{\max}) + G(\gamma). \tag{38}$$

Therefore, we have

$$\begin{aligned}
&\Pr\left\{f(T^{\max}) \leq f(W^{\max}) + G(\gamma) \middle| T^{\max}\right\} \\
\geq& \Pr\left\{\zeta T^{\max}\mathbb{1}_{\{T^{\max}\geq \widehat{J}(\gamma)/\zeta\}} \leq (A^{\max}+1)W^{\max}+1 \middle| T^{\max}\right\} \\
\geq&(1-\gamma)^2.
\end{aligned} \tag{39}$$

Hence, we have the desired result. □

Corollary 1. *For any $\gamma \in (0,1)$, we have*

$$\mathbb{E}\left[f(T^{\max}[t])\right] \leq \frac{1}{(1-\gamma)^2}\mathbb{E}\left[f(W^{\max}[t])\right] + G(\gamma), \tag{40}$$

where $G(\gamma)$ is some positive value depending on γ.

Proof. In the rest of the proof, we omit the time index associated with $T^{\max}[t]$ and $W^{\max}[t]$ due to space restrictions. Lemma 2 immediately implies

$$\begin{aligned}
&\mathbb{E}\left[f(W^{\max}) \middle| T^{\max}\right] \\
\geq&\mathbb{E}\left[f(W^{\max}) \middle| f(W^{\max}) \geq f(T^{\max}) - G(\gamma), T^{\max}\right](1-\gamma)^2 \\
\geq& \left(f(T^{\max}) - G(\gamma)\right)(1-\gamma)^2.
\end{aligned} \tag{41}$$

Hence, we have

$$f(T^{\max}) \le \frac{1}{(1-\gamma)^2}\mathbb{E}\left[f(W^{\max})\middle|T^{\max}\right] + G(\gamma), \tag{42}$$

By taking the expectation on both sides of the above inequality, we have the desired result. □

Having established these lemmas and corollary, we are ready to prove Proposition 1.

Proof. Consider the Lyapunov function

$$V(\mathbf{R},\mathbf{T}) \triangleq \sum_{i=1}^{I}\sum_{j\in\mathcal{N}_i} R_{i,j} f\left(T_{i,j}\right).$$

In the rest of the proof, due to the space restrictions, we omit the time index t associated with various variables and vectors, and use X^+ to denote the variable $X[t+1]$ without causing any confusion. Then, we have

$$\begin{aligned}\Delta V \triangleq& V(\mathbf{R}^+,\mathbf{T}^+) - V(\mathbf{R},\mathbf{T})\\ =&\sum_{i=1}^{I}\Big(\sum_{j\in\mathcal{N}_i^+} R^+_{i,j} f\left(T^+_{i,j}\right) - \sum_{j\in\mathcal{N}_i} R_{i,j} f\left(T_{i,j}\right)\Big).\end{aligned} \tag{43}$$

We first consider the term $\sum_{j\in\mathcal{N}_i^+} R^+_{i,j} f\left(T^+_{i,j}\right)$.

$$\begin{aligned}&\sum_{j\in\mathcal{N}_i^+} R^+_{i,j} f\left(T^+_{i,j}\right)\\ \overset{(a)}{=}&\sum_{j\in\mathcal{N}_i} R^+_{i,j} f\left(T^+_{i,j}\right) + \sum_{j\in\mathcal{A}_i} R^+_{i,j} f\left(T^+_{i,j}\right)\\ \overset{(b)}{=}&\sum_{j\in\mathcal{N}_i} f\left(\left(T_{i,j}+1\right)\left(1-\mathbb{1}_{\{R^+_{i,j}=0\}}\right)\right)R^+_{i,j} + f(1)\sum_{j\in\mathcal{A}_i}\left\lceil\frac{F_{i,j}}{c^{\max}}\right\rceil\\ \overset{(c)}{=}&\sum_{j\in\mathcal{N}_i} f\left(T_{i,j}+1\right)R^+_{i,j}\left(1-\mathbb{1}_{\{R^+_{i,j}=0\}}\right) + f(1)\sum_{j\in\mathcal{A}_i}\left\lceil\frac{F_{i,j}}{c^{\max}}\right\rceil\\ =&\sum_{j\in\mathcal{N}_i} f\left(T_{i,j}+1\right)R^+_{i,j} + f(1)\sum_{j\in\mathcal{A}_i}\left\lceil\frac{F_{i,j}}{c^{\max}}\right\rceil\\ \overset{(d)}{\le}&\sum_{j\in\mathcal{N}_i}\left(f\left(T_{i,j}\right)+f'\left(x_{i,j}\right)\right)\left(R_{i,j}-S_{i,j}\mathbb{1}_{\{C_{i,j}=c^{\max}\}}\right)\\ &+ f(1)\sum_{j\in\mathcal{A}_i}\left\lceil\frac{F_{i,j}}{c^{\max}}\right\rceil\\ \overset{(e)}{\le}&\sum_{j\in\mathcal{N}_i} R_{i,j} f\left(T_{i,j}\right) + \sum_{j\in\mathcal{N}_i} R_{i,j} f'\left(T_{i,j}\right)\\ &-\sum_{j\in\mathcal{N}_i} f\left(T_{i,j}\right)S_{i,j}\mathbb{1}_{\{C_{i,j}=c^{\max}\}} + f(1)\sum_{j\in\mathcal{A}_i}\left\lceil\frac{F_{i,j}}{c^{\max}}\right\rceil,\end{aligned} \tag{44}$$

where step (a) from the fact that flows at time $t+1$ are composed of existing flows and newly arriving flows at time t; (b) uses the dynamics of $T_{i,j}$ and the fact that the newly arriving flows are not served in the current slot; (c) follows from the assumption that $f(0)=0$; (d) uses the Mean Value Theorem for some $x_{i,j}$ between $T_{i,j}$ and $T_{i,j}+1$; (e) follows from the fact that $f'(y)$ is non-increasing and non-negative due to the function $f(y)$ being non-decreasing and concave for any $y\ge 0$.

By substituting (44) into (43), we have

$$\begin{aligned}\Delta V \le& f(1)\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i}\left\lceil\frac{F_{i,j}}{c^{\max}}\right\rceil + \sum_{i=1}^{I}\Big(\sum_{j\in\mathcal{N}_i} R_{i,j}[t] f'\left(T_{i,j}\right)\\ &-\sum_{j\in\mathcal{N}_i} f\left(T_{i,j}\right)S_{i,j}\mathbb{1}_{\{C_{i,j}=c^{\max}\}}\Big)\\ \le& f(1)\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_i}\left\lceil\frac{F_{i,j}}{c^{\max}}\right\rceil + \sum_{i=1}^{I}\Big(\sum_{j\in\mathcal{N}_i} R_{i,j} f'\left(T_{i,j}\right)\\ &-\sum_{j\in\mathcal{N}_i} f\left(W_{i,j}\right)S_{i,j}\mathbb{1}_{\{C_{i,j}=c^{\max}\}}\Big),\end{aligned} \tag{45}$$

where the last step uses the fact that TSLS value of each flow is always not greater than its corresponding age, i.e., $W_{i,j}\le T_{i,j}$ for any $j\in\mathcal{N}_i$ and $t\ge 0$. Thus, we have

$$\begin{aligned}\mathbb{E}\left[\Delta V\right] \le& f(1)\sum_{i=1}^{I}\rho_i + \sum_{i=1}^{I}\mathbb{E}\Big[\sum_{j\in\mathcal{N}_i} R_{i,j} f'\left(T_{i,j}\right)\\ &-\sum_{j\in\mathcal{N}_i} f\left(W_{i,j}\right)S_{i,j}\mathbb{1}_{\{C_{i,j}=c^{\max}\}}\Big].\end{aligned} \tag{46}$$

Next, we consider the term $\mathbb{E}\left[\sum_{j\in\mathcal{N}_i} R_{i,j} f'(T_{i,j})\right]$. In particular, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{j\in\mathcal{N}_i} R_{i,j} f'\left(T_{i,j}\right)\middle|W^{\max},T^{\max}\right]\\ \overset{(a)}{\le}&\mathbb{E}\Bigg[A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\sum_{\tau=t-T^{\max}}^{t-(1-\gamma)W^{\max}} f'(t-\tau)\\ &+\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\Bigg|W^{\max},T^{\max}\Bigg]\\ =&\mathbb{E}\Bigg[A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\sum_{m=(1-\gamma)W^{\max}}^{T^{\max}} f'(m)\\ &+\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\Bigg|W^{\max},T^{\max}\Bigg]\\ \overset{(b)}{\le}&A_i^{\max}\left\lceil\frac{F_i^{\max}}{c_i^{\max}}\right\rceil\left(f(T^{\max})-f((1-\gamma)W^{\max})+f'(1)\right)\\ &+\mathbb{E}\Bigg[\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\mathbb{1}_{\mathcal{E}_0}\Bigg|W^{\max},T^{\max}\Bigg]\\ &+\mathbb{E}\Bigg[\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\mathbb{1}_{\overline{\mathcal{E}_0}}\Bigg|W^{\max},T^{\max}\Bigg]\\ \overset{(c)}{\le}&A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\left(f(T^{\max})-f((1-\gamma)W^{\max})+f'(1)\right)\\ &+\mathbb{E}\Bigg[\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\mathbb{1}_{\mathcal{E}_0}\Bigg|W^{\max},T^{\max}\Bigg]\\ &+A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\left(f((1-\gamma)J(\gamma)-1)-f(1)+f'(1)\right),\end{aligned} \tag{47}$$

where step (a) follows from the fact that the amount of incoming workload in each time slot is not greater than $A_i^{\max}\lceil F_i^{\max}/c_i^{\max}\rceil$; (b) is true for $\mathcal{E}_0 \triangleq \{W^{\max} \geq J(\gamma)\}$, and both (b) and (c) use [8, Lemma 3.4] that for any $f \in \mathcal{G}$,

$$\sum_{m=M_L}^{M_U} f'(m) \leq f(M_U) - f(M_L) + f'(1), \tag{48}$$

where $M_U \geq M_L \geq 1$.

Next, we consider the second term in (47). For any $w \geq J(\gamma)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\middle| W^{\max}=w\right]\\
&\overset{(a)}{=}\mathbb{E}\left[\sum_{\tau=t-(1-\gamma)W^{\max}[t]+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\middle| \mathcal{E}_1, W^{\max}=w\right]\\
&\qquad\cdot \Pr\left\{\mathcal{E}_1 \middle| W^{\max}=w\right\} \quad + \quad \Pr\left\{\overline{\mathcal{E}}_1 \middle| W^{\max}=w\right\}\cdot\\
&\mathbb{E}\left[\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\sum_{j\in\mathcal{A}_i[\tau]} R_{i,j}[\tau]\middle| \overline{\mathcal{E}}_1, W^{\max}=w\right]\\
&\overset{(b)}{\leq}\left(\rho+\gamma A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\right)\sum_{\tau=t-(1-\gamma)W^{\max}+1}^{t-1} f'(t-\tau)\\
&=\left(\rho+\gamma A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\right)\sum_{m=1}^{(1-\gamma)W^{\max}-1} f'(m)\\
&\overset{(c)}{\leq}\left(\rho+\gamma A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\right)\left(f((1-\gamma)W^{\max}-1)-f(1)+f'(1)\right),
\end{aligned} \tag{49}$$

where step (a) is true for $\mathcal{E}_1 \triangleq \bigcap_{\tau\in[t-(1-\gamma)W^{\max},t-1]}\mathcal{F}_\tau$ and $\mathcal{F}_\tau$ being the event that all flows arriving after time $t-(1-\gamma)W^{\max}$ are not served at time slot τ; (b) uses Lemma 1; (c) uses (48).

By substituting (49) into (47), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j\in\mathcal{N}_i} R_{i,j} f'\left(T_{i,j}\right)\middle| W^{\max}, T^{\max}\right]\\
&\overset{(a)}{\leq} A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\left(f(T^{\max})-f((1-\gamma)W^{\max})\right)\\
&\quad+\left(\rho_i+\gamma A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\right)f((1-\gamma)W^{\max}-1)+B_{1,i}(\gamma)\\
&\overset{(b)}{\leq} A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\left(f(T^{\max})-f(W^{\max})\right)\\
&\quad+\left(\rho_i+\gamma A_i^{\max}\left\lceil\frac{F_i^{\max}}{c^{\max}}\right\rceil\right)f(W^{\max})+B_{2,i}(\gamma),
\end{aligned} \tag{50}$$

where step (a) is true for $B_{1,i}(\gamma) \triangleq f'(1)\big(\rho+(2+\gamma)A_i^{\max}\lceil F_i^{\max}/c^{\max}\rceil\big) + A_i^{\max}\lceil F_i^{\max}/c^{\max}\rceil f\left((1-\gamma)J(\gamma)-1\right)$; (b) uses the definition of $f \in \mathcal{G}$ and is true for $B_{2,i}(\gamma) \triangleq B_{1,i}(\gamma)+A_i^{\max}\lceil F_i^{\max}/c^{\max}\rceil G_0$ and G_0 is some positive constant.

Next, we focus on the term

$$\sum_{i=1}^{I}\mathbb{E}\left[\sum_{j\in\mathcal{N}_i} f(W_{i,j})S_{i,j}\mathbb{1}_{\{C_{i,j}=c^{\max}\}}\right].$$

$$\begin{aligned}
&\sum_{i=1}^{I}\mathbb{E}\left[\sum_{j\in\mathcal{N}_i} f(W_{i,j})S_{i,j}\mathbb{1}_{\{C_{i,j}=c^{\max}\}}\right]\\
&\overset{(a)}{\geq}\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-(1-p_I)^{Q(t-(1-\gamma)W^{\max},t)/\nu^{\max}}\right)\right]\\
&\overset{(b)}{\geq}\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-(1-p_I)^{Q(t-(1-\gamma)W^{\max},t)/\nu^{\max}}\right)\mathbb{1}_{\mathcal{E}_2}\right]\\
&\overset{(c)}{\geq}\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-(1-p_I)^{Q_0(W^{\max})/\nu^{\max}}\right)\mathbb{1}_{\mathcal{E}_2}\right]\\
&\overset{(d)}{=}\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-\frac{\gamma}{(W^{\max})^2}\right)\mathbb{1}_{\mathcal{E}_2}\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-\frac{\gamma}{(W^{\max})^2}\right)\mathbb{1}_{\mathcal{E}_2}\middle| W^{\max}\right]\right]\\
&=\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-\frac{\gamma}{(W^{\max})^2}\right)\Pr\left\{\mathcal{E}_2\middle| W^{\max}\right\}\right]\\
&\geq\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-\frac{\gamma}{(W^{\max})^2}\right)\mathbb{1}_{\{W^{\max}\geq J(\gamma)\}}\right.\\
&\qquad\left.\cdot\Pr\left\{\mathcal{E}_2\middle| W^{\max}\right\}\right]\\
&\overset{(e)}{\geq}\mathbb{E}\left[f((1-\gamma)W^{\max})\left(1-\frac{\gamma}{(W^{\max})^2}\right)^2\mathbb{1}_{\{W^{\max}\geq J(\gamma)\}}\right]\\
&\geq\left(1-\frac{\gamma}{(J(\gamma))^2}\right)^2\mathbb{E}\left[f((1-\gamma)W^{\max})\mathbb{1}_{\{W^{\max}\geq J(\gamma)\}}\right]\\
&\geq(1-2\gamma)\mathbb{E}\left[f((1-\gamma)W^{\max})\mathbb{1}_{\{W^{\max}\geq J(\gamma)\}}\right]\\
&=(1-2\gamma)\,\mathbb{E}\left[f((1-\gamma)W^{\max})\right]\\
&\quad-(1-2\gamma)\,\mathbb{E}\left[f(1-\gamma)W^{\max})\mathbb{1}_{\{W^{\max}<J(\gamma)\}}\right]\\
&\geq(1-2\gamma)\,\mathbb{E}\left[f((1-\gamma)W^{\max})\right]-(1-2\gamma)\,f((1-\gamma)J(\gamma))\\
&\overset{(f)}{\geq}(1-2\gamma)\,\mathbb{E}\left[f(W^{\max})\right]-B_3(\gamma),
\end{aligned} \tag{51}$$

where step (a) is true for $\nu^{\max} \triangleq \max_i\lceil F_i^{\max}/c^{\max}\rceil$, and follows from the definition of $Q(\tau,t)$ and the fact that if at least one flow in the set $\mathcal{N}(t-(1-\gamma)W^{\max},t)$ has the maximum channel rate, then $\sum_{i=1}^{K}\sum_{j\in\mathcal{N}_i} f(W_{i,j})S_{i,j}$ is at least $f((1-\gamma)W^{\max})$ according to our proposed TSLS-based scheduling algorithm; (b) is true for the event $\mathcal{E}_2 \triangleq \{Q(t-(1-\gamma)W_{\max},t) \geq Q_0(W_{\max})\}$; (c) simply uses the fact that $1-(1-p_I)^x$ is non-decreasing function of x; (d) follows from the definition of Q_0; (e) uses inequality (12); (f) is true for $B_3(\gamma) \triangleq (1-2\gamma)f((1-\gamma)J(\gamma)) + G_1$ and $G_1 > 0$ satisfies $f((1-\gamma)W^{\max}) \leq f(W^{\max}) + G_1$.

Hence, we have

$$\mathbb{E}\left[\Delta V\right] \overset{(a)}{\leq} \mathbb{E}\left[f(T^{\max}) - f(W^{\max})\right] \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil + \left(\sum_{i=1}^{I} \rho_i + \gamma \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil - (1-2\gamma) \right) \mathbb{E}[f(W_{\max})] + \sum_{i=1}^{K} B_{2,i}(\gamma) + B_3(\gamma), \tag{52}$$

Therefore, we have

$$\mathbb{E}\left[\Delta V\right] \overset{(a)}{\leq} \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil \left(\frac{1}{(1-\gamma)^2} - 1 \right) \mathbb{E}\left[f(W^{\max})\right] + \left(1 - 2\epsilon + \gamma \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil - (1-2\gamma) \right) \mathbb{E}[f(W^{\max})] + B_4(\gamma)$$
$$\overset{(b)}{\leq} \left(-2\epsilon + \gamma \left(1 + \frac{2-\gamma}{(1-\gamma)^2} \right) \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil + 2\gamma \right) \mathbb{E}[f(W^{\max})] + B_4(\gamma), \tag{53}$$

where step (a) uses Corollary 1 and is true for $B_4(\gamma) \triangleq \sum_{i=1}^{K} B_{2,i}(\gamma) + B_3(\gamma) + \sum_{i=1}^{K} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil G(\gamma)$. By selecting γ sufficiently small such that

$$\gamma \left(1 + \frac{2-\gamma}{(1-\gamma)^2} \right) \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil + 2\gamma \leq \epsilon, \tag{54}$$

we have

$$\mathbb{E}[\Delta V \leq -\epsilon \mathbb{E}[f(W^{\max})] + B_4(\gamma)$$
$$\leq -\epsilon(1-\gamma)^2 \mathbb{E}[f(T^{\max})] + B(\gamma), \tag{55}$$

where the last step uses Corollary 1 again and is true for $B(\gamma) \triangleq B_4(\gamma) + (1-\gamma)^2 G(\gamma)$.

By summing the inequality (55) over $t = 0, 1, 2, \ldots, M-1$, we have

$$\limsup_{M\to\infty} \frac{1}{M} \sum_{t=0}^{M-1} \mathbb{E}[f(T^{\max})] \leq \frac{B(\gamma)}{\delta}, \tag{56}$$

where $\delta \triangleq \epsilon(1-\gamma)^2$.

Since all flows at time slot t arrived at the system after time $t - T^{\max}$, we have

$$\sum_{i=1}^{K} \sum_{j \in \mathcal{N}_i} R_{i,j} \leq T^{\max} \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil, \tag{57}$$

which implies that for any $f \in \mathcal{G}$, we have

$$f\left(\sum_{i=1}^{I} \sum_{j \in \mathcal{N}_i} R_{i,j} \right) \overset{(a)}{\leq} f\left(T_{\max} \sum_{i=1}^{I} A_i^{\max} \left\lceil \frac{F_i^{\max}}{c^{\max}} \right\rceil \right) \overset{(b)}{\leq} f(T_{\max}) + G', \tag{58}$$

where step (a) uses the fact that $f \in \mathcal{G}$ is non-decreasing; (b) follows from the definition of $f \in \mathcal{G}$ and is true for some G'.

By substituting (58) into (56), we have

$$\limsup_{M\to\infty} \frac{1}{M} \sum_{t=0}^{M-1} \mathbb{E}\left[f\left(\sum_{i=1}^{I} \sum_{j \in \mathcal{N}_i[t]} R_{i,j}[t] \right) \right] \leq \frac{B(\gamma)}{\delta} + G' < \infty.$$

This implies stability-in-the-mean property and thus the underlying Markov Chain is positive recurrent [7]. □

7 CONCLUSION

In this work, we proposed a round-robin-like algorithm that has desired throughput, delay-insensitivity and short-term fairness performance in wireless dynamic fading networks, where flows dynamically arrive at the system and leave the system once their service requests are completed. We maintained a time-since-last-service (TSLS) counter for each flow, which keeps track of the time since the last service of the flow, and incorporated both TSLS and channel rate into the scheduling design, namely TSLS-based scheduling algorithm. We established the throughput-optimality property of our proposed algorithm, and demonstrate its delay-insensitivity and excellent short-term fairness performance in comparison to various existing policies through simulations.

REFERENCES

[1] Bonald, T., Borst, S., and Proutière, A. How mobility impacts the flow-level performance of wireless data systems. In *Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)* (Hong Kong, China, March 2004).

[2] Borst, S. Flow-level performance and user mobility in wireless data networks. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 366*, 1872 (2008), 2047–2058.

[3] Chen, Y., Wang, X., and Cai, L. Head-of-line access delay-based scheduling algorithm for flow-level dynamics. *IEEE Transactions on Vehicular Technology 66*, 6 (2016), 5387–5397.

[4] Dai, J. G. On the positive harris recurrence for multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability* (1995), 49–77.

[5] Demers, A., Keshav, S., and Shenker, S. Analysis and simulation of a fair queueing algorithm. In *ACM SIGCOMM Computer Communication Review* (1989), vol. 19, pp. 1–12.

[6] Ji, B., Joo, C., and Shroff, N. B. Delay-based back-pressure scheduling in multihop wireless networks. *IEEE/ACM Transactions on Networking 21*, 5 (2012), 1539–1552.

[7] Kumar, P., and Meyn, S. Stability of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control 40* (1995), 251–260.

[8] Li, B., Eryilmaz, A., and Srikant, R. Emulating round-robin in wireless networks. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing* (2017), ACM, p. 21.

[9] Lu, S., Bharghavan, V., and Srikant, R. Fair scheduling in wireless packet networks. In *ACM SIGCOMM* (1997), vol. 27, ACM, pp. 63–74.

[10] N.Walton. Concave switching in single and multihop networks. *Queueing Systems 81*, 2 (2015), 265–299.

[11] Rhee, I., Warrier, A., Min, J., and Xu, L. DRAND: Distributed randomized TDMA scheduling for wireless ad hoc networks. *IEEE Transactions on Mobile Computing 8*, 10 (2009), 1384–1396.

[12] Rybko, A. N., and Stolyar, A. L. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problemy Peredachi Informatsii 28*, 3 (1992), 3–26.

[13] Sadiq, B., and de Veciana, G. Throughput optimality of delay-driven maxweight scheduler for a wireless system with flow dynamics. In *Proc. Allerton Conference on Communications, Control and Computing (Allerton)* (Monticello, IL, October 2009).

[14] Tassiulas, L., and Ephremides, A. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control 36*, 12 (1992), 1936–1948.

[15] Tassiulas, L., and Ephremides, A. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory 39* (1993), 466–478.

[16] van de Ven, P., Borst, S., and Shneer, S. Instability of maxweight scheduling algorithms. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)* (Rio de Janeiro, Brazil, April 2009).