On Scheduling Policies With Heavy-Tailed Dynamics in Wireless Queueing Systems

SHENGBO CHEN\textsuperscript{1}, (Member, IEEE), LANXUE ZHANG\textsuperscript{2}, (Student Member, IEEE), CONG SHEN\textsuperscript{2}, (Senior Member, IEEE), KEPING YU\textsuperscript{3,4}, (Member, IEEE), SAN HLAING MYINT\textsuperscript{3}, AND ZHENG WEN\textsuperscript{5}

\textsuperscript{1}School of Computer and Information Engineering, Henan University, Kaifeng 475001, China
\textsuperscript{2}Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904, USA
\textsuperscript{3}Global Information and Telecommunication Institute, Waseda University, Tokyo 169-8050, Japan
\textsuperscript{4}School of Fundamental Science and Engineering, Waseda University, Tokyo 169-8050, Japan
\textsuperscript{5}Shenzhen Boyi Technology Company Ltd., Shenzhen 518125, China

Corresponding author: Keping Yu (keping.yu@aoni.waseda.jp)

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant JP18K18044.

ABSTRACT This paper takes a system view and studies a wireless queueing system where heavy-tailness may occur both at the traffic arrival and in the form of the multi-user interference. With the rapid development of AI technologies, this heavy-tailed traffic model has become more prevalent in the current network system, such as the file or data size used in the deep learning algorithm. We first revisit the standard asymmetric queueing system with a mix of heavy-tailed and light-tailed traffic, but under a new variable-rate service model that not only better models the dynamics of the wireless medium but also includes the previous models as special cases. We then focus on the scheduling problem when heavy-tailed interference disrupts the serving link. The performance of queueing policies is investigated during an ON/OFF renewal channel process with heavy-tailed OFF periods, and the expected queue length and the throughput characteristic is studied under the priority as well as max-weight scheduling policies. The results show that the expected queue length of the heavy queue cannot be maintained as finite even under the most favorable priority policy. On the other hand, a priority policy can guarantee the finiteness of an expected queue length for the light queue, but the system is not throughput optimal any longer. It is further shown that no benefit can be provided by the max-weight scheduling policy to the light queue for the queue length behavior in a steady-state, though the system is always throughput optimal.

INDEX TERMS Heavy-tailed interference, queueing analysis, scheduling, artificial intelligence.

I. INTRODUCTION Scheduling policy is an important problem in a networking area, and we have seen lots of literature for variant systems, including wireless [1], ad-hoc [2], and multi-hop networks [3]. In this field, simple traffic models were considered in early research, such as Poisson or Markov-modulated processes, as well as simple service models. However, the network traffic nowadays has begun to show strong correlations and statistical similarity [4]. As these simple traffic models are not enough [5], heavy-tailed distributions have been adopted to model such traffic in networking area. In these packetized networks, the file sizes vary quite significantly, and data traffic is shown to be more bursty. In these packetized networks, the file sizes vary quite significantly, and data traffic is shown to be more bursty. For example, with the rapid development of AI technologly [6]–[9], the data size used in the deep learning algorithm might be heavy-tailed [10].

A majority of the existing scheduling policies analyses, when heavy-tailed traffic exists, focuses on asymmetric queueing models, which shows how the system behavior is influenced by heavy-tailed traffic. In these models, there is generally a mix of heavy-tailed and light-tailed traffic [11]–[13], which can describe the heterogeneous traffic types in the network. Regarding the performance metric, two criteria are often considered: expected queue length and...
throughput optimality. Throughput is a widely used performance metric, since it shows the best performance region that a scheduling policy can achieve. Second, since the delay is proportional to the queue length, a finite expected queue length in the steady-state is preferred whenever possible.

The authors [13] investigate a single-server system with two parallel wireline queues, where one queue has heavy-tailed traffic (“heavy queue”) while the other has light-tailed traffic (“light queue”). They show that no scheduling policy can guarantee the heavy queue’s queue length asymptotic, and the light queue may act as poorly as the heavy queue if an inappropriate policy is selected, including the famous max-weight policy [14]. Based on this work, [15] obtains the queue length distribution in a steady-state, and characterizes its tail coefficient for both the max-weight-α and log-max-weight policies, which prefers to provide more service to the light queue compared to the max-weight policy. The authors in [14] prove that max-weight is throughput optimal. The authors in [16] study the delay stability of the max-weight policies for more complex networks. They propose a novel approach that combines the fluid model with either renewal theory (for instability results) or stochastic Lyapunov theory (for stability results) in [17]. Recently, the authors [18] have extended their work to multi-hop networks, and show that the back-pressure-α policy can stabilize the light flows delay while the original back-pressure policy cannot, which is similar to the result for the max-weight policies in the single-hop case.

In the wireless queueing problems, channel dynamics have to be considered. The authors in [19] introduce an ON/OFF channel between the server and the queue, and study both the max-weight-α and log-max-weight policies. They characterize the tail behavior of the queue length in steady-state under heavy-tailed arrival processes. The same ON/OFF channel model is also considered in [20], where they study the impact of heavy-tailed traffic on the response time. Both inter-queue and intra-queue scheduling choices are considered to optimize the delay, and the analysis shows that the response time tail of the light queue is sensitive to the intra-queue policy choice.

All of the aforementioned works have laid the theoretical foundation for asymmetric queueing models with both heavy-tailed and light-tailed traffic. However, they may not be enough when we consider a wireless queueing system. This is because some unique characteristics for a wireless system have not been fully considered in previous studies. First, the literature only considers channels that are either always on [13], [15], [16], or can be modeled by some light-tailed ON/OFF channels, e.g., the time-varying ON/OFF channels with independent and identically distributed (i.i.d.) Bernoulli processes [19]. Such channel modeling assumes that the channel only experiences short time correlations, which might not be true in reality. A better capture for the wireless dynamics, such as large-scale and small-scale channel fading [21] needs to be considered.

Second, the heavy-tailed distributions are only considered in the traffic model. However, for better understanding the influence of heavy-tailed distributions in a wireless system, we should not only consider the traffic arrival, but also the potential heavy-tailed interference from other users. This is because of the distinctive property of wireless communications, that is, the broadcasting nature [21]. As a result, if the data transmission for one user shows some form of heavy-tailed behavior [22], it will result in a heavy-tailed interference to other users in the system when the interference is strong.

In this paper, we perform a systematic study on the impact of heavy-tailed dynamics to the scheduling policies in a wireless queueing system. First we introduce a general variable-rate service model that can be used to study the standard asymmetric queueing system with a mix of heavy-tailed and light-tailed distributions for either traffic arrival or multi-user interference. This new model incorporates the effect of channel fading and multi-user interference, and thus better captures the dynamics of the wireless medium. We first analyze the model with heavy-tailed traffic arrivals and general time-varying service rate, and study the expected queue-length behavior in a steady-state of priority, max-weight, and max-service-rate scheduling policies. We then focus on the scheduling problem in the presence of heavy-tailed interference, which can be modeled by the proposed variable-rate service model with a heavy-tailed OFF period to one queue (“heavy queue”) and light-tailed for the other (“light queue”).

It is shown that the queue length for the heavy queue is unstable even under the most favorable priority-for-H policy. Thus, no scheduling policies can guarantee the heavy queue’s queue-length stability. Regarding the light queue, we show that the queue-length distribution is light-tailed under the priority-for-L policy, but throughput optimal does not exist any more. For the max-weight scheduling policy, a threshold-base performance for the light queue is discovered. If the arrival rate for the light queue is below some threshold, any scheduling policy can stabilize the light queue. On the contrary, when the arrival rate to the light queue is greater than the threshold, the expected queue length in a steady-state is proven to be infinite under the max-weight scheduling policy. Hence, no benefit can be provided by the max-weight scheduling to the light queue in terms of the steady-state queue-length performance, although the system is throughput optimal.

To the best of the authors’ knowledge, no result has been obtained for a wireless network that contains both heavy-tailed and light-tailed interference. From a methodological point of view, the main technical difficulty for this model comes from the correlation of interference across time slots, which makes the queue-length behavior no longer a Markov process. As we will see, the technicality renders the standard tools, such as Lyapunov functions,

1 Only priority and max-weight policies are studied for the interference model, as the max-service-rate policy is not applicable.
difficult to use. We thus resort to the creation of artificial queues that can provide delay bounds to the original model, and use an asymptotic argument to derive the steady-state results.

The rest of the paper is organized as follows. Section II shows the system model and some preliminaries. The impact of mixing heavy-tailed and light-tailed traffic arrival is studied with a general variable-rate service model in Section III. We then analyze the performance of priority and max-weight scheduling policies with heavy-tailed multi-user interference in Section IV. We conclude the paper in Section V, with a discussion for future directions.

II. SYSTEM MODEL AND PRELIMINARIES

A. DEFINITIONS

We first show some relevant definitions, which will be extensively used in the remainder of this paper.

Definition 1 (Heavy-Tailness): A nonnegative random variable \( X \) is called heavy-tailed if \( \mathbb{E}[X^2] = \infty \). It is light-tailed otherwise.

We note that there are different definitions of heavy-tailness in the literature. For example, light-tailed is sometimes defined if the random variable is of exponential type, and heavy-tailed otherwise. Our definition is the same as [16], [23].

Definition 2 (Tailed Coefficient): The tailed coefficient \( C \) of a nonnegative random variable \( X \) is defined as \( C = \min\{C \in \mathbb{R}^+ | \mathbb{E}[X^C] = \infty \} \).

Definition 3 (Rate Stability): A queue \( Q(t) \) is stable if

\[
\lim_{t \to \infty} \frac{Q(t)}{t} = 0 \text{ with probability 1.}
\]

Based on the queueing theorem, a queue is stable if and only if the mean service rate is greater than the mean arrival rate.

Definition 4 (Capacity Region): The capacity region of a queueing system with \( K \) queues is the rate region where the mean rate tuple \((\lambda_1, \ldots, \lambda_K)\) is stably supportable, i.e., the queues can be stable for some scheduling policy. Here \( \lambda_i \) denotes the mean arrival rate for queue \( i \), \( i \in \{1, \ldots, K\} \).

Definition 5 (Throughput Optimality): We call a scheduling policy throughput optimal if the queues can be stabilized when the mean rate tuple is within the capacity region.

Definition 6 (Stable-State Queue Length): A queue \( Q(t) \) is stable and its stable-state queue length is defined as

\[
q = \lim_{t \to \infty} \mathbb{E}[Q(t)].
\]

B. THE WIRELESS QUEUEING SYSTEM MODEL

In this section, we introduce the general service-rate queueing model. We consider a time-slotted wireless system with a single server and two parallel queues \( H \) and \( L \), as illustrated in Figure 1. The server can serve at most one queue at a time and the service rate is a normalized 1 packet per time slot.\(^2\)

![Figure 1](image_url)

Both queues are assumed to have infinite buffer. We denote the queue lengths as \( Q_H(t) \) and \( Q_L(t) \), and their arrival processes \( A_H(t) \) and \( A_L(t) \), respectively, where \( A_i(t) \) is the number of packets arriving at queue \( i \) in time slot \( t \), \( i \in \{H, L\} \). We further use \( q_H \) and \( q_L \) to denote the steady-state queue length when they exist. We assume that the arrival process \( A_i(t) \) is stationary with mean \( \lambda_i \) and tail coefficient \( C_i \), \( i \in \{H, L\} \). All arrivals are assumed to be i.i.d. over time and of each queue, and occur at the end of each time slot.

We will refer to queue \( H \) and \( L \) as the “heavy queue” and “light queue”, respectively, although the heavy-tailed and light-tailed distributions may refer to either traffic arrival or channel OFF period, depending on the context.

Both queues are connected to the server via wireless time-varying channels, which fluctuate over time according to some random processes. In practice, this fluctuation may be due to channel noise or other interference in the wireless medium. Specifically, we denote the wireless channel between queue \( i \) and server at time \( t \) as \( G_i(t) \), which is stationary with mean \( g_i \), \( i \in \{H, L\} \). Correspondingly, the available service rate, which is determined by a combination of the service rate and the channel state, is a time-varying random process and denoted as \( S_i(t) \) for \( i \in \{H, L\} \). We assume that \( S_i(t) \) is stationary, has mean \( s_i \), and has support \( \mathcal{K}_i \), \( i \in \{H, L\} \). The variable service rate processes \( S_H(t) \) and \( S_L(t) \) are independent of the arrival processes and of each other. We assume that the channel conditions, or equivalently the service rates, are known by the queues before the scheduling decision. Note that the channel states may not be i.i.d. over time, which is particularly suitable when the interference is persistent. This is a case that will be considered in Section IV.

III. HEAVY-TAILED TRAFFIC ARRIVALS

We first analyze the general variable-rate service model of Section II with a mix of heavy-tailed and light-tailed traffic arrival distributions. To that end, we further assume that the arrival processes to the \( L \) (\( H \)) queue, \( A_L(t) \) (\( A_H(t) \)), is light-tailed (heavy-tailed) according to Definition 1. When the scheduler chooses to serve queue \( i \), the service rate is a random process \( S_i(t) \), which is i.i.d. over time and across queues. We assume that both \( S_H(t) \) and \( S_L(t) \) are light-tailed,

\(^2\)The service rate can be a random process in our model. Since we also consider the random channel variations, such service rate randomness can be absorbed into the channel variation, while keeping a normalized packet rate.
i.i.d. over time, and distributed according to some discrete distributions with \( P[S_L(\cdot) = k] = p_k, k \in \mathcal{R}_L, \) and \( P[S_H(\cdot) = j] = r_j, j \in \mathcal{R}_H. \) Obviously, \( s_L = \sum_{k \in \mathcal{R}_L} p_k k, \) \( s_H = \sum_{j \in \mathcal{R}_H} r_j j. \)

For this model, the capacity region, according to Definition 4, can be easily derived and is given by

\[
\Lambda = \bigcup_{a_{kj} \in A} \{(\lambda_H, \lambda_L)|0 < \lambda_L < \sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} a_{kj} p_k r_j k, 0 < \lambda_H < \sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} (1 - a_{kj}) p_k r_j j\},
\]

where

\[ A \triangleq \{a_{kj} : a_{k0} = 1, a_{0j} = 0, 0 \leq a_{kj} \leq 1, \forall k \in \mathcal{R}_L, j \in \mathcal{R}_H\}. \tag{2} \]

The boundary of the capacity region consists of a piecewise linear curve. Let \((\Lambda_L, \Lambda_H)\) denote a boundary point of the capacity region, then we have that \(0 < \Lambda_L < s_L.\) Moreover, for a given \(\Lambda_L,\) we can determine the corresponding \(\Lambda_H\) by solving the following linear programming problem:

\[
\begin{align*}
\text{maximize} & \quad \sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} (1 - a_{kj}) p_k r_j j \\
\text{subject to} & \quad \sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} a_{kj} p_k r_j k = \Lambda_L \\
& \quad a_{k0} = 1, \quad \forall k \in \mathcal{R}_L \\
& \quad a_{0j} = 0, \quad \forall j \in \mathcal{R}_H \\
& \quad 0 \leq a_{kj} \leq 1, \quad \forall k \in \mathcal{R}_L, j \in \mathcal{R}_H. \tag{3}
\end{align*}
\]

A pictorial illustration of the capacity region is shown as the gray area in Fig. 2.

**A. PRIORITY FOR H**

In this policy, the scheduler selects queue \(H\) as long as \(S_H(t) Q_H(t) > 0,\) and serves queue \(L\) in other scenarios. Note that the policy does not need to know either the actual queue length or the service rate, only that they are nonzero. Again, this represents the best-case scenario for the heavy queue, and the steady-state queue length analysis will shed light into the behavior of the heavy queue. We have the following result.

**Proposition 1**: Under the priority-for-H scheduling policy, the heavy queue is stable and \(q_H = \infty.\)

**Proof**: It is equivalent to proving

\[
\lim_{t \to \infty} E[Q_H(t)] = \infty.
\]

The main idea behind the proof is to consider the renewal intervals that commence at the beginning of each busy period of queue \(H.\) Define the renewal reward process \(R(t) = Q_H(t).\) We have

\[
\lim_{t \to \infty} E[R(t)] = \frac{E[R]}{E[T]},
\]

from the key renewal theorem, where \(E[R]\) denotes the expected reward accumulated over a renewal interval, and \(E[T] < \infty\) is the mean renewal interval. It is therefore enough to show that

\[
E[\sum_{i=0}^{T} Q_H(t)] = \infty. \tag{4}
\]

To prove Eqn. (4), let us condition on the busy period commencing at time 0 with a burst of size \(b\) to queue \(H.\) After this instant, assuming for the sake of a lower bound that there are no further bursts arriving at queue \(H,\) then with high probability, queue \(H\) drains at rate \(s_H.\) So the reward is at least \(O(b)\) for \(O(b)\) time slots. Thus, for some constant \(K,\)

\[
E[\sum_{i=0}^{T} Q_H(t)] \geq E[KB \cdot b] = E[KB^2] = \infty,
\]

where the last expectation is infinite because the initial burst size has tail coefficient \(C_H\) that is smaller than 2. \(\square\)

**B. PRIORITY FOR L**

The priority-for-L policy is similarly defined as Section III-A. We have the following main result.

**Proposition 2**: Under the priority-for-L scheduling policy, the following statements hold:

1. If \(\lambda_H > (1 - \lambda_L / s_L + p_0 \lambda_L / s_L) s_H,\) the heavy queue is unstable, and no steady-state exists.
2. If \(\lambda_H < (1 - \lambda_L / s_L + p_0 \lambda_L / s_L) s_H,\) the heavy queue is stable, and \(q_H = \infty.\)
3. The light queue is stable and \(q_L < \infty.\)

**Proof**: For the light queue, we can prove \(q_L < \infty\) by considering the Lyapunov function

\[
V_L(Q_L(t)) = \frac{1}{2} Q_L^2(t).
\]

Expanding the term \(E[V_L(Q_L(t + 1))|Q_L(t)]\) at \(Q_L(t),\) we have:

\[
e E[V_L(Q_L(t + 1))|Q_L(t)] = E[\frac{1}{2} (Q_L(t) + A_L(t) - \min(Q_L(t), S_L(t)))^2|Q_L(t)]
\]
Therefore, the average service rate of
In this case, the heavy queue is unstable, and no steady-state
which means that there must exist a rate pair \((\lambda_L, \lambda_H)\) that satisfies
\(\sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} a_{kj} \rho_k r_j k = \lambda_L\),
from the capacity region (1). By choosing \(a_{0j} = 0\), \(a_{1j} = 1\), and \(a_{ij} = \frac{\lambda_L - \lambda_H r_H}{(1 - \rho_L r_H)}\), we know that \(\sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} a_{ij} \rho_k r_j k = \lambda_L\)
is satisfied. Also, since \(\frac{\lambda_L - \lambda_H r_H}{(1 - \rho_L r_H)} < \frac{\lambda_L}{s_L}\), we have
\[
\Lambda_H \geq \sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} (1 - \frac{\lambda_L - \lambda_H r_H}{(1 - \rho_L r_H)}) p_k r_j j
\]
\[
> \sum_{k \in \mathcal{R}_L} \sum_{j \in \mathcal{R}_H} \frac{\lambda_L}{s_L} p_k r_j j
\]
\[
\geq (1 - \frac{\lambda_L}{s_L} + p_0 \frac{\lambda_L}{s_L}) s_H,
\]
which means that there must exist a rate pair \(\{\lambda_L, \lambda_H\}\) in the capacity region that satisfies \(\lambda_H > (1 - \frac{\lambda_L}{s_L} + p_0 \frac{\lambda_L}{s_L}) s_H\). In this case, the heavy queue is unstable, and no steady-state
exists. As a result, the priority-for-\(L\) policy is not throughput optimal.
If \(\lambda_H < (1 - \frac{\lambda_L}{s_L} + p_0 \frac{\lambda_L}{s_L}) s_H\), then \(q_H\) is stable. However, the expected queue length of queue \(H\) is infinite even under the priority-for-\(H\) policy, so it will remain infinite under the priority-for-\(L\) policy.

**Proposition 3:** Under any nonidling scheduling policy, if \(\lambda_L < s_L r_H\), then \(q_L < \infty\).

**Proof:** Suppose queue \(L\) is only served when event \(\{S_H(t) = 0\}\) occurs. Since \(S_H(\cdot)\) and \(S_L(\cdot)\) are both i.i.d. and \(S_L(\cdot)\) is light-tailed, we can derive that the service rate of queue \(L\) (denoted as \(S_L(\cdot)\) in every time slot is light-tailed with mean \(s_L r_0\)). Precisely, \(P[S_L(\cdot) = 0] = 1 - r_0 + r_0 p_0\), \(P[S_L(\cdot) = k] = r_0 p_k\) for \(k \in \mathcal{R}_L\). Therefore, as in Proposition 2, if \(\lambda_L < s_L r_0\), we can similarly prove that \(E[q_L] < \infty\). Moreover, the service opportunity is always offered to queue \(L\) when \(S_H(t) = 0\), no matter what policy it is. Therefore, if \(\lambda_L < s_L r_0\), \(E[q_L] < \infty\) holds under any nonidling scheduling policy.

\(\square\)

**C. MAX WEIGHT**
The max-weight policy, unlike the priority policies, makes use of the queue length information at each time slot. Hence, we assume that the scheduler knows \(Q_H(t)\) and \(Q_L(t)\) in addition to the service rate information \(S_H(t)\) and \(S_L(t)\). More precisely, the scheduler compares
\[
S_H(t) Q_H(t) \geq S_L(t) Q_L(t)
\]
and chooses to serve the queue that wins the competition.

For the max-weight policy, the following result states that the steady-state queue length will be infinite for large \(\lambda_L\).

**Proposition 4:** Under max-weight scheduling with a bounded service rate \(S_L(\cdot) \leq S_{L_{\text{max}}}\) for queue \(L\), if \(\lambda_L > s_L r_0\), then \(E[Q_L] = \infty\).

**Proof:** Conditioning on the busy period commencing with a burst of size \(b\) at the heavy queue, \(Q_H(t) > S_{L_{\text{max}}} Q_L(t)\) is almost surely satisfied. Therefore, the heavy queue will drain at a rate of at most \(s_H b\) with high probability, and the light queue will build up at a rate of \(\lambda_L - s_L r_0\) with high probability. In order to get more service than \(s_L r_0\), queue \(L\) will at least build up to \(O(b)\) level. After that, the two queues drain together, with most of the slots being used to serve queue \(H\). Since queue \(H\) drains at most by \(O(b)\), queue \(L\) stays at \(O(b)\) level for at least \(O(b)\) time slots. Therefore, we have
\[
E[\sum_{i=0}^{T} Q_L(i)] \geq E[K b \cdot b] = E[K b^2] = \infty,
\]
where \(T\) is the renewal interval of queue \(L\)’s busy period. Similar to the proof of Proposition 1, we have that \(E[Q_L] = \infty\).

\(\square\)

**D. MAX SERVICE RATE**
The max-service-rate policy is similar to the multi-user diversity scheme [24], [25] and is often used in wireless communication systems. It chooses to serve the nonempty queue.
that has the highest instantaneous service rate. Specifically, the scheduler compares $S_H(t) \cdot 1_{\{Q_H(t) > 0\}} \geq S_L(t) \cdot 1_{\{Q_L(t) > 0\}}$ and serves the queue that wins the competition.

We characterize the behavior of both queues in Proposition 5. Interestingly, the stability results for both queues exhibit some threshold effects.

**Proposition 5:** Under the max-service-rate policy,

1. queue $H$ is heavy-tailed with tail coefficient $C_H = 1$ if $\lambda_H < \mu_H$, and unstable otherwise;
2. queue $L$ is light-tailed if $\lambda_L < \mu_L$, and unstable otherwise.

**Proof:** For queue $H$, the average service rate is

$$\mu_H = \mathbb{E}[S_H | S_H \geq S_L] + P(Q_H(t) = 0)\mathbb{E}[S_H | S_H < S_L]$$

where $a_0 = 0, a_k = 1, a_j = P(Q_H(t) = 0)1_{k \leq j} + P(Q_L(t) > 0)1_{k > j}$. Therefore, similar as in the proof of Proposition 1, we can derive that if $\lambda_H < \mu_H$, queue $H$ is stable and has an infinite expected queue length. If $\lambda_H > \mu_H$, queue $H$ is unstable and no steady-state exists.

For queue $L$, the average service rate is

$$\mu_L = \mathbb{E}[S_L | S_L \geq S_L] + P(Q_H(t) = 0)\mathbb{E}[S_H | S_H < S_L]$$

where $a_0 = 0, a_k = 1, a_j = P(Q_H(t) = 0)1_{k \leq j} + P(Q_L(t) > 0)1_{k > j}$. Therefore, similar to the proof of Proposition 2, we can derive that if $\lambda_L < \mu_L$, queue $L$ has a finite expected queue length. However, if $\lambda_L > \mu_L$, queue $L$ is unstable and no steady-state exists.

Obviously, $\{\mu_L, \mu_H\}$ is in the capacity region. Hence, if $\Lambda_L = \mu_L$, then $\Lambda_H \geq \mu_H$. Therefore, if $\Lambda_L < \mu_L$, then $\Lambda_H > \mu_H$. As a result, the max-service-rate policy is not throughput optimal.

Note that this is a general service-rate queueing model, which not only better models the wireless channel dynamics but also includes the previous ON/OFF models as special cases. For the scheduling policies, similar properties hold for both models.

**IV. HEAVY-TAILED MULTI-USER INTERFERENCE**

**A. A MOTIVATIONAL EXAMPLE**

A heterogeneous wireless cellular network is considered, where femtocells are placed overlaid with macrocells. For femtocells, such deployment allows them to share the spectrum with the macrocell and thus the spectrum utilization can be improved. A typical scenario is depicted in Figure 3, where both the macrocell and femtocell have their respective users. There is a femto user FU1, which is in the interference range of the macrocell, while another femto user FU2 is not. Hence, if heavy-tailed traffic is transmitted on the downlink in the macrocell, such as the log-normal type of distributions for call durations as reported in [26], FU1 will experience heavy-tailed interference. While FU2 does not suffer from such interference, but may be influenced by some out-of-cell interference, which is light-tailed.

Denoting the packet processing capability of FU$n$ as $V_n(t)$ at time $t$, the femtocell needs to decide how to send each FU’s packets for them to process, with knowledge of $V_n(t)$, and/of $Q_n(t), n \in \{H, L\}$. As we can see, this problem can be analyzed using the general variable service rate model.

**B. ASSUMPTIONS AND PRELIMINARY ANALYSIS**

We consider the variable service rate model described in Section II-B, but only highlight the new assumptions for the heavy-tailed interference problem in this section. We consider both arrival processes $A_H$ and $A_L$ to be light-tailed, as the heavy-tailness is reflected in the channel process instead of the arrival. The $G_n$ is modeled as an ON/OFF renewal process $\{O_n, U_n\}$, with alternating independent ON period $O_n$ and OFF period $U_n$, for $n \in \{H, L\}$. The process $O_H, U_L, O_L$ are light-tailed. However, $U_H$ is heavy-tailed because of the heavy-tailed interference. We provide a summary of these parameters in Table 1.

![Figure 3. A motivational example of mix of both heavy-tailed and light-tailed interference in a heterogeneous cellular network.](image-url)

**TABLE 1. Summary of parameters.**

| Parameters | Expected Value | Variance |
|------------|----------------|----------|
| $A_H$      | $\lambda_H$    | $\phi_H$ |
| $A_L$      | $\lambda_L$    | $\phi_L$ |
| $O_H$      | $a_1$          | $b_1$    |
| $U_H$      | $a_2$          | $\infty(b_2)$ |
| $O_L$      | $a_3$          | $b_3$    |
| $U_L$      | $a_4$          | $b_4$    |

A more general approach would be to combine heavy-tailed traffic arrival with heavy-tailed interference. However, such approach would result in various combinations of the mixture of heavy-tailed and light-tailed arrival and interference, which is beyond the scope of this paper.
focus on the case in which the rate pair is in the capacity region.

In this section, we investigate the scheduling performance of the heavy-tailed interference system under different policies. We focus on both priority (for queue $H$ and $L$) and max-weight policies in the analysis. The max-service-rate policy becomes trivial because of the ON/OFF nature of the channel and the constant service rate. Also, it is clear that a scheduling policy is only useful in determining which queue should be selected. Furthermore, no queue can be served when both channels are OFF.

We should emphasize that our model is different from the existing literature. The main difference is that the heavy-tailed distribution applies to the OFF period of queue $H$’s channel, instead of the Markovian requirement. As a result, some key differences will be brought about by the *heavy-tailed channel dynamics* even if we apply the same scheduling policy in the new model. Methodologically, *Lyapunov functions* are often adopted to deal with this type of problems. However, this method requires the queue length to serve when both $G_H$ and $G_L$ are ON. As only one channel is ON and the other is OFF, the queue with the ON channel should be selected. Furthermore, no queue can be served when both channels are OFF.

We should emphasize that our model is different from the existing literature. The main difference is that the heavy-tailed distribution applies to the OFF period of queue $H$’s channel, instead of the Markovian requirement. As a result, some key differences will be brought about by the *heavy-tailed channel dynamics* even if we apply the same scheduling policy in the new model. Methodologically, *Lyapunov functions* are often adopted to deal with this type of problems. However, this method requires the queue length to serve when both $G_H$ and $G_L$ are ON. As only one channel is ON and the other is OFF, the queue with the ON channel should be selected. Furthermore, no queue can be served when both channels are OFF.

We consider the priority-for-$L$ policy, where queue $L$ is served as long as it is nonempty and channel $G_L$ is ON. The result for the expected queue length and throughput optimality is shown as follows.

**Proposition 6:** Under the priority-for-$L$ policy, the expected queue length of queue $L$ in steady-state is finite. However, when $\lambda_H > p_H (1 - \lambda_L)$, the system is not throughput optimal.

**Proof:** We will first prove that queue $L$ is queue-length stable. For each OFF period $U_L$ of queue $L$, the cumulative arrival is $A_{UL} = \sum_{i=1}^{N} A_L(i)$. The expected value of $A_{UL}$ is derived by:

$$E[A_{UL}] = E[E[A_{UL} | U_L = N]]$$
$$= E[E[\sum_{i=1}^{N} A_L(i) | U_L = N]]$$
$$= E[N \lambda_L]$$
$$= a_4 \lambda_L.$$

The variance of $A_{UL}$ can be derived as:

$$\text{Var}[A_{UL}] = E[A_{UL}^2] - (E[A_{UL}])^2$$
$$= E[E[A_{UL}^2 | U_L = N]] - (E[A_{UL}])^2$$
$$= E[E[\sum_{i=1}^{N} A_L(i)^2 | U_L = N]] - a_4^2 \lambda_L^2$$
$$= E[N (\phi_L + \lambda_L^2) + N(N-1) \lambda_L^2] - a_4^2 \lambda_L^2$$
$$= E[N \phi_L + N^2 \lambda_L^2] - a_4^2 \lambda_L^2$$
$$= a_4 \phi_L + (a_4^2 + b_4) \lambda_L^2 - a_4^2 \lambda_L^2$$
$$= a_4 \phi_L + \lambda_L^2 b_4.$$
and $A_{UL}$ at the first time slot for each ON period. Denote $A_{UL}(t) = A_{UL}(t) \cdot I(t)$ (t is the first time slot of an $O_L$ period), then the arrival process for queue $L_1$ can be formulated as

$$A_{L_1}(t) = A_L(t) + A_{UL}(t).$$

Further, we notice that all the ON periods form a renewal process, where the process $A_{UL}$ is as a renewal reward process associated. From renewal theory [27], the average arrival rate for $A_{UL}$ is provided by following:

$$\lim_{T \to \infty} \left( \frac{1}{T} \int A_{UL} \cdot I(t) \text{ is the first time slot of an } O_L \text{ period} \, dt \right) = \frac{E[A_{UL}]}{E[O_L]} = \frac{a_1 \lambda_L}{a_3}.$$

Therefore, we obtain the total arrival rate for queue $L_1$, given by $\lambda_L + a_1 \lambda_L / a_3 = \lambda_L(a_3 + a_4)/a_3 < 1$, where the last holds because $\lambda_L / a_3 < a_1 / a_3 + a_2$. Hence, we conclude that queue $L_1$ is stable, which means that queue $L$ is stable.

Next, we will show that the expected queue length of queue $L$ in steady-state is finite. A second artificial queue system $L_2$ with two subqueues $L_{21}$ and $L_{22}$ is considered, as depicted in Figure 6. The arrival for queue $L_{21}$ is $A_L$ and the server’s service rate is $a_1$. The other queue $L_{22}$ has the arrival $A_{UL}$ at the service rate $a_3$. We know that the queue length of $L_1$ will be smaller than the sum of the queue length of $L_{21}$ and $L_{22}$, since some service will be wasted if either queue $L_{21}$ or $L_{22}$ is empty. Similarly, we consider another artificial queue system $L_{23}$ with the arrival process $A_{UL}(t)$ and the service rate $a_4 / a_3 + a_4$. The queue length performance of queue $L_{22}$ in an $O_L$ period should be identical as that of queue $L_{23}$ in any time slot. Therefore, we know that the expected queue length of queue $L_{22}$ and $L_{23}$ should be the same. The expected queue length of queue $L_{21}$ and $L_{23}$, which are discrete-time Geo/G/1 queues, is provided by [28]:

$$E[D] = \frac{\lambda^2 b_2 - \lambda \rho + \lambda_2 b}{2(1 - \rho)}, \quad (8)$$

where $\lambda$ and $\lambda_2$ are the first and second moment of the arrival rate, $b$ and $b_2$ are the first and second moment of the service rate for each arrival, and $\rho$ is the traffic intensity. We can see that the expected queue lengths are finite under both queue $L_{21}$ and $L_{23}$, according to Eqn. (8) and Little’s Law. Hence, we conclude that the expected queue length of queue $L$ is finite.

Finally, we will prove that the system is no longer throughput optimal when $\lambda_H > p_H(1 - \lambda_L)$. For queue $H$, it can only get service under two conditions: 1) when $G_H$ is ON and $G_L$ is OFF; or 2) when both $G_H$ and $G_L$ are ON and queue $L$ is empty. Since queue $L$ is queue length stable under the priority-for-$L$ policy, by applying Little’s law, we know that the fraction of time that queue $L$ is busy is equal to $\frac{\lambda_L}{p_L}$. As a result, the average service rate of queue $H$ is

$$\mu_H = p_H(1 - \rho_L) + p_H p_L (1 - \frac{\lambda_L}{p_L}) = p_H(1 - \lambda_L).$$

If $\lambda_H > p_H(1 - \lambda_L)$, the average service rate is smaller than the average arrival rate for queue $H$. Thus queue $H$ will not be stable even if the mean rate ($\lambda_H$, $\lambda_L$) is in the capacity region. Therefore, we conclude that this scheduling policy is not throughput optimal.

Here we show some simulation results in Fig. 7 and 8. We randomly generate the traffic arrival $A_L$, $A_H$, channel ON period $O_L$, $O_H$ and channel OFF period $U_L$ following the heavy-tailed distribution and channel OFF period $U_H$ following the heavy-tailed Pareto distribution. Assume that the expectations are satisfied $a_1 = a_2 = a_3 = a_4 = 3$. Figure 7 shows the queue length change in $Q_L$ with $\lambda_L$ under...
the priority-for-\(L\) for \(\lambda_H = 0.1\) and \(\lambda_H = 0.5\), respectively. When \(\lambda_L < 0.5\), the queue length of queue \(L\) is finite on both condition. Once \(\lambda_L > 0.5\), \(Q_L\) becomes infinite because \(\lambda_L > p_H = a_3/(a_3 + a_4) = 3/(3+3) = 0.5\). Besides, we plot the queue length change in \(Q_H\) with \(\lambda_H\) under the priority-for-\(L\) for \(\lambda_L = 0.1\) and \(\lambda_L = 0.5\), respectively, in Figure 8. When \(\lambda_L = 0.1\), the queue length of queue \(H\) is finite on condition of \(\lambda_H < p_H(1 - \lambda_L) = 0.5 \times (1 - 0.1) = 0.45\). In case of \(\lambda_H > 0.45\), the arrival rate is greater than the average service rate and \(Q_H\) becomes unstable. The condition of \(\lambda_L = 0.5\) is similar to \(\lambda_L = 0.1\) elaborated above.

**D. PRIORITY FOR H POLICY**

For the priority-for-\(H\) policy, the server will serve queue \(H\) whenever it is nonempty and channel \(G_H\) is ON. Proposition 7 shows a negative result for the expected queue length under this policy.

**Proposition 7:** For the priority-for-\(H\) scheduling policy, the expected queue length of queue \(H\) in steady-state is infinite.

**Proof:** Similar to the previous deduction, for each OFF period \(U_H\) of queue \(H\), the cumulative arrival during this period is given by \(A_{U_H} = \sum_{i=1}^{N} A_H(i)\). We can obtain the mean and variance of \(A_{U_H}\) as:

\[
\mathbb{E}[A_{U_H}] = a_2 \lambda_H,
\]

\[
\text{Var}[A_{U_H}] = a_2 \phi_H + \lambda_H^2 b_2 = \infty,
\]

respectively. As no packet is allowed to be transmitted during the OFF period, the queue behavior of queue \(H\) will be the identical as a new artificial queue \(H_1\), where the time consists of the ON periods of queue \(H\), and remove all the OFF periods. The number of arrivals in the first slot of an ON period \(O_H\) is equal to the sum of all arrivals in the previous OFF period and the arrivals of the current time slot. We denote \(A_{H_1}(t) = A_{U_H}(t) \cdot 1[t \text{ is the first time slot of an } O_H \text{ period}].\) We have the arrival process of the artificial queue \(H_1\) as

\[
A_H(t) = A_H(t) + A_{H_1}(t).
\]

Similarly, we consider two artificial queue: \(H_2\) with the arrival process \(A_{H_2}\) and the service rate \(1\) packet per time slot; and \(H_3\) with the arrival process \(A_{U_H}\) and the service rate \(O_H\) packets per time slot. The queue length performance of queue \(H_2\) in an \(O_H\) period should be identical as the queue length performance of queue \(H_3\) in any time slot. Therefore, the expected queue length of queue \(H_2\) should be the same as \(H_3\). Since \(\text{Var}[A_{U_H}]\) is infinite, the expected queue length of queue \(H_3\) is infinite. As a result, the expected queue length of queue \(H_2\) is also infinite according to Eqn. (8). It is clear that the expected queue length of queue \(H_1\) is greater than the expected queue length of queue \(H_2\), which leads to the conclusion.

In the system considered above, the priority-for-\(H\) policy is the best one that prefers the heavy queue. We thus claim that the expected queue length of queue \(H\) is always infinite under any scheduling policy.

**E. MAX-WEIGHT POLICY**

For the max-weight scheduling policy, the server compares the length of \(Q_H(t)\) and \(Q_L(t)\) when both channels are ON, and serves the longer queue. This subsection is to analyze the queue length performance under the max-weight scheduling policy. Because of a technical reason, we assume that the process \(O_H\) is always less than some constant \(O_{H_{\text{Inf}}}\).

**Proposition 8:** For the max-weight scheduling policy, the expected queue length of queue \(L\) in steady-state is:

1. finite, if \(\lambda_L < (1 - p_H)p_L\);
2. infinite, if \(\lambda_L > (1 - p_H)p_L\).

**Proof:**

There are four subcases of the channel conditions as follows:

1. \(P_1 \triangleq (1 - p_H)(1 - p_L),\) if both \(G_H\) and \(G_L\) are OFF, denoted as case \(A_1\);
2. \(P_2 \triangleq (1 - p_H)p_L,\) if \(G_H\) is OFF and \(G_L\) is ON, denoted as case \(A_2\);
3. \(P_3 \triangleq p_H(1 - p_L),\) if \(G_H\) is ON and \(G_L\) is OFF, denoted as case \(A_3\); and
4. \(P_4 \triangleq p_H p_L,\) if both \(G_H\) and \(G_L\) are ON, denoted as case \(A_4\).

First, we will show that the expected queue length of queue \(L\) is finite only if \(\lambda_L < (1 - p_H)p_L\). This is done by proving that the expected queue length of queue \(L\) in steady-state is finite when queue \(L\) get served at \(A_2\), that is, the conclusion holds no matter what scheduling policy is used.

Suppose queue \(L\) can only be served at \(A_2\). Define \(U\) as the time duration between two successive \(A_2\)’s. As queue \(L\) cannot be served during \(U\), we know that \(O_L\) occurs only in the \(O_H\) period, rather than in the \(U_H\) period. An illustration of \(U\) is shown in Figure 9.

Clearly, for each \(O_L\) that occurs during \(U\), we have

\[
O_L(i) < O_{H_{\text{Inf}}}.
\]

Since each \(O_L(i)\) is i.i.d., we have

\[
P[O_L(i) \leq O_{H_{\text{Inf}}}] = P[O_L \leq O_{H_{\text{Inf}}}] \triangleq p_0.
\]

We use \(N\) to denote the number of time slots that \(O_L\) occurs during \(U\), then we can obtain that

\[
P[N = n] \leq P[O_L(1) \leq O_{H_{\text{Inf}}}, \cdots, O_L(n) \leq O_{H_{\text{Inf}}}] = p_0^n.
\]
\[
\begin{align*}
&= \prod_{i=1}^{n} \mathbb{P}[O_L(i) \leq O_{Hw}] \\
&= p_0^n.
\end{align*}
\]

Thus we have
\[
\begin{align*}
E[N] \leq \sum_{n=0}^{\infty} np_0^n = \frac{p_0}{(1 - p_0)^2}, \\
E[N^2] \leq \sum_{n=0}^{\infty} n^2p_0^n = \frac{p_0(p_0 + 1)}{(1 - p_0)^3}.
\end{align*}
\]

We can show that
\[
0 < U \leq \sum_{j=1}^{N} O_L(j) + \sum_{k=1}^{N+1} U_L(k).
\]

Hence, the mean and variance of \( U \) satisfy
\[
\begin{align*}
&= \mathbb{E}[U] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{N} O_L(k) + \sum_{k=1}^{N+1} U_L(k)\right] \mathbb{P}[N = n] \\
&= \mathbb{E}[a_1 \cdot n + a_4 \cdot (n + 1)] \\
&= a_4 + (a_3 + a_4)\mathbb{E}[N] \\
&\leq a_4 + (a_3 + a_4) \cdot \frac{p_0}{(1 - p_0)^2} < \infty,
\end{align*}
\]

and
\[
\begin{align*}
\sigma^2 &\triangleq \text{Var}[U] \\
&\geq \mathbb{E}[U^2] - \mathbb{E}[U]^2 \\
&\leq \mathbb{E}\left[\sum_{k=1}^{N} O_L(k) + \sum_{k=1}^{N+1} U_L(k)\right]^2 \mathbb{P}[N = n] - m^2 \\
&= \mathbb{E}[n(b_3 + a_3^2) + n(n - 1)a_3^2 + (n + 1)(b_4 + a_4^2) \\
&+ (n + 1)n_3^2 + 2n(n + 1)a_3a_4] - m^2 \\
&= \mathbb{E}[n^2(a_3^2 + a_3^2 + 2a_3a_4) + n(b_3 + b_4 + 4a_4^2) \\
&+ 2a_3a_4] \mathbb{E}[N] + a_3^2 + b_4 - m^2 \\
&\leq (a_3^2 + a_3^2 + 2a_3a_4) \mathbb{E}[N] + a_3^2 + b_4 - m^2 \\
&\leq a_3^2 + 4a_3^2 + 2a_3a_4 \frac{p_0(p_0 + 1)}{(1 - p_0)^3} + (b_3 + b_4 + 2a_4^2) \\
&+ 2a_3a_4 \frac{p_0}{(1 - p_0)^2} + a_3^2 + b_4 - m^2 \\
&< \infty.
\end{align*}
\]

In each \( U \), we have the cumulative arrival is \( A_U = \sum_{i=1}^{U} A_L(i) \). Similar to the proof of Proposition 6, we can obtain the mean and variance of \( A_U \) as
\[
\begin{align*}
\mathbb{E}[A_U] &= m\lambda_L, \\
\text{Var}[A_U] &= m\phi_L + \lambda_L^2 \sigma^2.
\end{align*}
\]

Because packets can only be transmitted during \( A_2 \), the queue behavior of queue \( L \) will be identical to a new artificial queue \( L_3 \), where the time line only consists of the \( A_2 \) periods without all other periods. Denote \( A_{U2}(t) = A_U(t) \cdot 1 \) (\( t \) is the first time slot of an \( A_2 \) period). The arrival process of queue \( L_3 \) can be written as
\[
A_{L3}(t) = A_L(t) + A_{U2}(t).
\]

Also we notice that all packets arriving at queue \( L \) arrive at queue \( L_3 \), with the only exception that they arrive in shorter periods. Let \( T_2 \) denote as the time period of \( A_2 \) and \( T_0 \) as the time period between the first slot of \( A_2 \) and the last slot before the next \( A_2 \). Then we have
\[
\sum_{i=1}^{T_2} A_{L3}(i) = \sum_{i=1}^{T_0} A_L(i).
\]

Since \( P_2 = \mathbb{E}[T_2/T_0] \), we have
\[
\mathbb{E}[A_{L3}] = \mathbb{E}[A_L]/P_2 = \frac{\lambda_L}{P_2} < 1
\]

This proves that queue \( A_{L3} \) is stable. We also have
\[
\mathbb{E}[A_{U2}] = \mathbb{E}[A_{L3}] - \mathbb{E}[A_L] = \frac{\lambda_L}{P_2} - \lambda_L < 1 - P_2.
\]

Next, we will prove that the expected queue length of queue \( L_3 \) is finite. We consider a second artificial queue system \( L_4 \) with two subqueues \( L_{41} \) and \( L_{42} \). The arrival process of queue \( L_{41} \) is \( A_2 \) and the server’s service rate is \( P_2 \). The other queue \( L_{42} \) serves the arrival \( A_{U2} \) and the rate is \( 1 - P_2 \). We also consider another artificial queue \( L_{43} \) with the arrival process \( A_L \) and the service rate \( A_2 \) packets per slot. The queue length performance of queue \( L_{42} \) in an \( A_2 \) period and the queue length of queue \( L_{43} \) in any time slot should be the same. Hence, we know that the expected queue length of queue \( L_{42} \) and \( L_{43} \) should be identical.

The expected queueing delays of queue \( L_{41} \) and \( L_{43} \) are both finite according to Eqn. (8). Combined with Little’s Law, we conclude that the expected queue lengths of queue \( L_{41} \) and \( L_{43} \) are finite. Note that the actual queue length of queue \( L_3 \) will be smaller than the summation of queue \( L_{41} \) and \( L_{42} \), since some service will be wasted if either queue \( L_{41} \) or \( L_{42} \) is empty. Hence, the expected queue length of queue \( L_3 \) is finite, and as a result the expected queue length of queue \( L \) is finite.

Then we will study the case \( \lambda_L > (1 - p_H)p_L \). In this subcase, queue \( L \) has to contend for another \( \lambda_L - P_2 \) fraction of time from sub-case 4 in order that the average service rate is greater than the average arrival rate. Now we consider a packet arrives at queue \( L \) at time slot \( \tau \), the queueing delay of this packet is denoted as \( W_L(\tau) \). We notice that \( W_L(\tau) \) consists of four parts, \( T_1, T_2, T_3, T_4 \), denoting the number of slots for the four sub-cases. We consider the condition \( Q_H(\tau) > Q_L(\tau) \). The scenario with best delay performance occurs when no packet arrives at queue \( H \) after time slot \( \tau \). The scheduling policy keeps selecting queue \( H \) until both queue lengths become equal, and then queue \( L \) is scheduled...
when it is available. We denote \( T \) as the least time such that \( Q_L(\tau + T) > Q_H(\tau + T) \). Then, we have
\[
(W_L(\tau - T)I_{[t \in (\tau, t + T)]} \geq Q_L(\tau),
\forall t \in (T, T, \cdots, T + W_L(\tau)).
\]
By taking expectation on both sides conditioned on the event \( \Gamma = \{Q_H(\tau) > Q_L(\tau)\} \), we have
\[
E[(W_L(\tau - T))|\Gamma|P(t \in A_2 \cup A_4) \geq E[Q_L(\tau)|\Gamma],
\]
which means that
\[
E[(W_L(\tau - T))|\Gamma| \geq \frac{1}{P_2 + P_4}E[Q_L(\tau)|\Gamma].
\]
(9)
Notice that \( T \) must satisfy the following:
\[
Q_L(\tau) + \sum_{k=0}^{T} A_L(\tau + k) - T_2 \geq Q_H(\tau) - T_3 - T_4.
\]
Again taking the conditional expectation on both sides, and together with the ergodic property of \( T_2, T_3, T_4 \), we have
\[
E[\sum_{k=0}^{T} A_L(\tau + k) + (P_3 + P_4 - P_2)T|\Gamma]
\geq E[Q_H(\tau) - Q_L(\tau)|\Gamma].
\]
Because \( T \) is a stopping time, according to the Wald’s equation, we have
\[
E[T|\Gamma] \geq \frac{1}{\lambda_L + P_3 + P_4 - P_2}E[Q_H(\tau) - Q_L(\tau)|\Gamma].
\]
(10)
Substituting (10) into (9), we have
\[
E[W_L(\tau)|\Gamma] \geq \frac{1}{\lambda_L + P_3 + P_4 - P_2}E[Q_H(\tau) - Q_L(\tau)|\Gamma]
+ \frac{1}{P_2 + P_4}E[Q_L(\tau)|\Gamma]
\geq \frac{1}{\lambda_L + P_3 + P_4 - P_2}E[Q_H(\tau)|\Gamma] + \left( \frac{1}{P_2 + P_4} - \frac{1}{\lambda_L + P_3 + P_4 - P_2} \right)E[Q_L(\tau)|\Gamma]
\geq \frac{1}{\lambda_L + P_3 + P_4 - P_2}E[Q_H(\tau)|\Gamma] + \left( \frac{1}{P_2 + P_4} - \frac{1}{P_3 + P_4} \right)E[Q_L(\tau)|\Gamma].
\]
If
\[
\frac{1}{P_2 + P_4} - \frac{1}{P_3 + P_4} > 0,
\]
we have
\[
E[W_L(\tau)|\Gamma] > \frac{1}{P_3 + P_4}E[Q_H(\tau)|\Gamma] = \infty,
\]
(11)
which concludes that the expected queue length of queue \( L \) is infinite based on Little’s Law.

Otherwise,
\[
\left( \frac{1}{P_3 + P_4} - \frac{1}{P_2 + P_4} \right)E[Q_L(\tau)|\Gamma] + E[W_L(\tau)|\Gamma]
\geq \frac{1}{P_3 + P_4}E[Q_H(\tau)|\Gamma].
\]
(12)
Again from Little’s Law, we have
\[
E[W_L(\tau)] = \frac{1}{\lambda_L}E[Q_L(\tau)].
\]
Now, if \( E[Q_L(\tau)] \) is infinite, we obtain the conclusion directly. Otherwise, \( E[Q_L(\tau)|\Gamma] \) is finite if \( E[Q_L(\tau)] \) is finite. We denote \( V \) as an upper bound for \( E[Q_L(\tau)|\Gamma] \). We can rewrite (12) as
\[
E[W_L(\tau)|\Gamma]
\geq \frac{1}{P_3 + P_4}E[Q_H(\tau)|\Gamma] - \left( \frac{1}{P_3 + P_4} - \frac{1}{P_2 + P_4} \right)V
\geq E[Q_H(\tau)|\Gamma] - \left( \frac{1}{P_3 + P_4} - \frac{1}{P_2 + P_4} \right)V.
\]
(13)
Next, we consider the event \( \Gamma^c = \{Q_H(\tau) \leq Q_L(\tau)\} \). It follows that
\[
E[W_L(\tau)|\Gamma^c] \geq E[Q_H(\tau)|\Gamma^c]
\geq E[Q_H(\tau)|\Gamma^c] - \left( \frac{1}{P_3 + P_4} - \frac{1}{P_2 + P_4} \right)V.
\]
(14)
Combining (13) and (14), we have
\[
E[W_L(\tau)] \geq E[Q_H(\tau)] - \left( \frac{1}{P_3 + P_4} - \frac{1}{P_2 + P_4} \right)V.
\]
(15)
From Proposition 6, we have \( E[Q_H(\tau)] \) is infinite. Therefore, the expected queue length of queue \( L \) is infinite by Little’s Law and (15), when \( \lambda_L > (1 - p_H)p_L \). This concludes the proof.

V. CONCLUSION

The scheduling policies are studied in a wireless system with both heavy-tailed and light-tailed dynamics, which may come from either traffic arrivals or multi-user interference. For heavy-tailed traffic arrivals, we studied a more general variable-rate service model, in which channel dynamics (as opposed to simple ON/OFF) are incorporated. We then focused on the more challenging heavy-tailed multi-user interference problem, and found that the heavy queue has an unstable queue length asymptotic, which cannot be guaranteed by any scheduling policies. Regarding the light queue, we proved that the priority-for-L policy can stabilize its queue length, though not throughput optimal any more. Max-weight scheduling is throughput optimal but cannot provide the queue length stability for the light queue. Furthermore, there exists a threshold effect for the light queue. When the traffic arrival rate for the light queue is less than the threshold, its queue length will be light-tailed under any scheduling policy. Neither the priority nor max-weight policy provides a satisfying queueing performance for the considered model with heavy-tailed interference. Some prior study [19] has
shown that revised max-weight policies, including log-max-weight and max-weight-\(\alpha\), can guarantee better performance for heavy-tailed traffic arrivals. We will investigate these scheduling policies under heavy-tailed traffic interference in the future work.

ACKNOWLEDGMENT
This article was presented in part at the 2017 Information Theory and Applications (ITA) Workshop, San Diego, CA, USA, February 2017 [29].

REFERENCES

[1] M. J. Neely, “Order optimal delay for opportunistic scheduling in multiuser wireless uplinks and downlinks,” IEEE/ACM Trans. Netw., vol. 16, no. 5, pp. 1188–1199, Oct. 2008.
[2] V. J. Venkataramanan, X. Lin, L. Ying, and S. Shakkottai, “On scheduling for minimizing end-to-end buffer usage over multiphop wireless networks,” in Proc. IEEE INFOCOM, Mar. 2010, pp. 1–9.
[3] L. Huang and M. J. Neely, “Delay efficient scheduling via redundant constraints in multihop networks,” Perform. Eval., vol. 68, no. 8, pp. 670–689, Aug. 2011.
[4] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, “On the self-similar nature of Ethernet traffic (extended version),” IEEE/ACM Trans. Netw., vol. 2, no. 1, pp. 1–15, Feb. 1994.
[5] V. Paxson and S. Floyd, “Wide area traffic: The failure of Poisson modeling,” IEEE/ACM Trans. Netw., vol. 3, no. 3, pp. 226–244, Jun. 1995.
[6] G. Gui, H. Huang, Y. Song, and H. Sari, “Deep learning for an effective nonorthogonal multiple access scheme,” IEEE Trans. Veh. Technol., vol. 67, no. 9, pp. 8440–8450, Sep. 2018.
[7] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao, “Flight delay prediction based on aviation big data and machine learning,” IEEE Trans. Veh. Technol., vol. 69, no. 1, pp. 140–150, Jan. 2020.
[8] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, “Fast beamforming design via deep learning,” IEEE Trans. Veh. Technol., vol. 69, no. 1, pp. 1065–1069, Jan. 2020.
[9] Y. Wang, M. Liu, J. Yang, and G. Gui, “Data-driven deep learning for automatic modulation recognition in cognitive radios,” IEEE Trans. Veh. Technol., vol. 68, no. 4, pp. 4074–4077, Apr. 2019.
[10] Y. Wang, D. K. Ramanan, and M. Hebert, “Learning to model the tail,” in Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS), Dec. 2017, pp. 7029–7039.
[11] Z. Shao and U. Madhow, “Scheduling heavy-tailed data traffic over the wireless Internet,” in Proc. IEEE 56th Veh. Technol. Conf., vol. 2, Jun. 2002, pp. 1158–1162.
[12] S. Borst, M. Mandjes, and M. van Uitert, “Generalized processor sharing with light-tailed and heavy-tailed input,” IEEE/ACM Trans. Netw., vol. 11, no. 5, pp. 821–834, Oct. 2003.
[13] M. G. Markakis, E. H. Modiano, and J. N. Tsitsiklis, “Scheduling policies for single-hop networks with heavy-tailed traffic,” in Proc. 47th Annu. Allerton Conf. Commun., Control, Comput. (Allerton), Sep. 2009, pp. 112–120.
[14] L. Tassiulas and A. Ephremides, “Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks,” IEEE Trans. Autom. Control, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.
[15] K. Jaganathan, M. Markakis, E. Modiano, and J. N. Tsitsiklis, “Queueing-length asymptotics for generalized max-weight scheduling in the presence of heavy-tailed traffic,” IEEE/ACM Trans. Netw., vol. 20, no. 4, pp. 1096–1111, Aug. 2012.
[16] M. G. Markakis, E. Modiano, and J. N. Tsitsiklis, “Max-weight scheduling in queueing networks with heavy-tailed traffic,” IEEE/ACM Trans. Netw., vol. 22, no. 1, pp. 257–270, Feb. 2014.
[17] M. G. Markakis, E. Modiano, and J. N. Tsitsiklis, “Delay analysis of the max-weight policy under heavy-tailed traffic via fluid approximations,” Math. Oper. Res., vol. 43, no. 2, pp. 460–483, May 2018.
[18] M. G. Markakis, E. Modiano, and J. N. Tsitsiklis, “Delay stability of back-pressure policies in the presence of heavy-tailed traffic,” IEEE/ACM Trans. Netw., vol. 24, no. 4, pp. 2046–2059, Aug. 2016.
[19] K. Jaganathan, M. G. Markakis, E. Modiano, and J. N. Tsitsiklis, “Throughput optimal scheduling over time-varying channels in the presence of heavy-tailed traffic,” IEEE Trans. Inf. Theory, vol. 60, no. 5, pp. 2896–2909, May 2014.
[20] J. Nair, K. Jaganathan, and A. Wierman, “When heavy-tailed and light-tailed flows compete: The response time tail under generalized max-weight scheduling,” IEEE/ACM Trans. Netw., vol. 24, no. 2, pp. 982–995, Apr. 2016.
[21] D. Tse and P. Viswanath, Fundamentals of Wireless Communication. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[22] S. Teymori and W. Zhuang, “Queue analysis and multiplexing of heavy-tailed traffic in wireless packet data networks,” Mobile Netw. Appl., vol. 12, no. 1, pp. 31–41, Jan. 2007.
[23] K. Park and W. Willinger, “Self-similar network traffic: An overview,” in Self-Similar Network Traffic and Performance Evaluation, K. Park and W. Willinger, Eds. Hoboken, NJ, USA: Wiley, 2000.
[24] R. Knopp and P. A. Humblet, “Information capacity and power control in single-cell multiuser communications,” in Proc. IEEE Inf. Conf. Commun., vol. 1, Jun. 1995, pp. 331–335.
[25] P. Viswanath, D. N. C. Tse, and R. Laroia, “Opportunistic beamforming using dumb antennas,” IEEE Trans. Inf. Theory, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.
[26] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, “Primary user behavior in cellular networks and implications for dynamic spectrum access,” IEEE Commun. Mag., vol. 47, no. 3, pp. 88–95, Mar. 2009.
[27] R. Gallager, Discrete Stochastic Processes. New York, NY, USA: Academic, 1996.
[28] H. Takagi, Queueing Analysis: Discrete-Time Systems, vol. 3. Amsterdam, The Netherlands: North-Holland, 1993.
[29] H. Wu, C. Shen, and S. Chen, “On scheduling policies in the presence of heavy-tailed interference,” in Proc. Inf. Theory Appl. Workshop (ITA), Feb. 2017, pp. 1–7.

SHENBO CHEN (Member, IEEE) received the B.E. and M.E. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2006 and 2008, respectively, and the Ph.D. degree from Ohio State University, Columbus, OH, USA, in 2013. From 2013 to 2019, he was a Senior Research Engineer with the Qualcomm Research Center, San Diego, CA, USA. He is currently a Professor with the School of Computer and Information Engineering, Henan University, China. He holds more than 50 U.S. patents on 5G and AI area. Dr. Chen received the Best Student Paper Award for WoSP 2013.

LANXUE ZHANG (Student Member, IEEE) is currently pursuing the bachelor’s degree in computer science and technology with the School of Computer and Information Engineering, Henan University, China.

CONG SHEN (Senior Member, IEEE) received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2002 and 2004, respectively, and the Ph.D. degree from the Electrical Engineering Department, UCLA, in 2009. From 2009 to 2014, he was with Qualcomm Research, San Diego, CA, USA. From 2015 to 2019, he was a specially-appointed Professor with the School of Information Science and Technology, University of Science and Technology of China (USTC). He is currently an Assistant Professor with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia. His research interests span a number of interdisciplinary areas in wireless communications, networking, and machine learning. He also serves as the Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE WIRELESS COMMUNICATIONS LETTERS.
KEPING YU (Member, IEEE) received the M.E. and Ph.D. degrees from the Graduate School of Global Information and Telecommunication Studies, Waseda University, Tokyo, Japan, in 2012 and 2016, respectively.

He was a Research Associate with the Global Information and Telecommunication Institute, Waseda University, from 2015 to 2019. He is currently a Junior Researcher with the Global Information and Telecommunication Institute, Waseda University. He has hosted and participated in a lot of research projects, including the Ministry of Internal Affairs and Communication (MIC) of Japan, the Ministry of Economy, Trade and Industry (METI) of Japan, the Japan Society for the Promotion of Science (JSPS), the Advanced Telecommunications Research Institute International (ATR) of Japan, the Keihin Electric Railway Corporation of Japan, and the Maspro Denkoh Corporation of Japan. He is also the Leader and a coauthor of the comprehensive book Design and Implementation of Information-Centric Networking (Cambridge University Press, 2020). He was involved in many standardization activities organized by ITU-T and ICNRG of IRTF, and contributed to the ITU-T Standards ITU-T Y.3071: Data Aware Networking (Information Centric Networking) Requirements and Capabilities and Y.3033-Data Aware Networking-Scenarios and Use Cases. His research interests include smart grids, information-centric networking, the Internet of Things, blockchain, and information security.

Dr. Yu had experience with editorial and conference organizations. He has served as a General Co-Chair and a Publicity Co-Chair of IEEE VTC2020-Spring EBTSRA workshop, a TPC Co-Chair of SCML2020, and a Session Chair of ITU Kaleidoscope 2016. Moreover, he has served as a TPC Member for the ITU Kaleidoscope 2020, IEEE VTC2020-Spring, IEEE CCNC 2020, IEEE WCNC 2020, IEEE VTC2019-Spring, ITU Kaleidoscope 2019, IEEE HotICN 2019, IEEE ICCC 2019, IEEE WPMC 2019, EEI 2019, and ICTIVE 2019. He is also the Editor of the IEEE Open Journal of Vehicular Technology (OJVT).

SAN HLAING MYINT received the B.C.Tech., B.C.Tech. (Hons.), and M.C.Tech. degrees from the University of Computer Studies at Mandalay, Myanmar, in 2004, 2005, and 2008, respectively, and the Ph.D. degree from Waseda University, Tokyo, Japan, in 2019. He was a Tutor with the Ministry of Science and Technology, Myanmar, from 2007 to 2015. From 2009 to 2014, he was with the University of Computer Studies at Yangon, Myanmar, where he performed research focused on the sustainable cloud computing. He is currently a Research Associate with the Global Information and Telecommunication Institute, Waseda University. His current researches focus on fifth-generation wireless networks and the analysis of channel coding methods.

ZHENG WEN received the B.E. degree in computer science and technology from Wuhan University, China, in 2009, and the M.Sc. and Ph.D. degrees from Waseda University, Tokyo, Japan, in 2015 and 2019, respectively. He became a Research Associate at Waseda University, in 2018, where he is currently an Assistant Professor (Lecturer) with the Department of Communication and Computer Engineering. His research interests include ICN/CCN for next-generation communication systems, AI, and the IoT.