MIT Open Access Articles

Throughput Optimal Scheduling Over Time-Varying Channels in the Presence of Heavy-Tailed Traffic

Massachusetts Institute of Technology

Throughput Optimal Scheduling over Time-Varying Channels in the presence of Heavy-Tailed Traffic

Krishna Jagannathan, Mihalis G. Markakis, Eytan Modiano, and John N. Tsitsiklis

Abstract

We study the problem of scheduling over time-varying links in a network that serves both heavy-tailed and light-tailed traffic. We consider a system consisting of two parallel queues, served by a single server. One of the queues receives heavy-tailed traffic (the "heavy queue"), and the other receives light-tailed traffic (the "light queue"). The queues are connected to the server through time-varying ON/OFF links, which model fading wireless channels. We first show that the policy that gives complete priority to the light-tailed traffic guarantees the best possible tail behavior of both queue backlog distributions, whenever the queues are stable. However, the priority policy is not throughput maximizing, and can cause undesirable instability effects in the heavy queue. Next, we study the class of throughput optimal max-weight-α scheduling policies. We discover a threshold phenomenon, and show that the steady-state light queue backlog distribution is heavy-tailed for arrival rates above a threshold value, and light-tailed otherwise. We also obtain the exact 'tail coefficient' of the light queue backlog distribution under max-weight-α scheduling. Finally, we analyze a log-max-weight (LMW) scheduling policy, and show that in addition to being throughput optimal, the LMW policy ensures that the light queue backlog distribution is light-tailed.

Index Terms

Scheduling, Queuing Analysis, Heavy-tailed traffic

I. INTRODUCTION

Scheduling conflicting communication links is an important task that arises in a variety of settings, including wireless networks and high speed switches. There is a large literature on scheduling conflicting links in a constrained queueing network, and many of these papers are based on the maximum-weight scheduling framework proposed in [20], [21]. The importance of maximum-weight scheduling is due to its 'throughput optimality' property. That is, it can stably support the largest set of traffic rates that is supportable by a given queueing network. For this reason, the max-weight family of scheduling policies has received much attention in various networking contexts, including switches [13], satellites [16], wireless [17], and optical networks [2].

Although throughput is an important first-order performance metric, a more discerning metric is the delay experienced by the traffic flows. While the throughput optimality of max-weight scheduling and its variants have been well understood for a while, the delay properties of these scheduling policies are not as thoroughly understood. Average delay bounds, such as those in [10], [17] can be derived using Lyapunov drift techniques; however, these are quite loose in general.

Existing results on the delay performance of max-weight policies indicate that these policies tend to perform well when the competing traffic sources are, loosely speaking, symmetric and well-behaved. This is intuitively due to the tendency of max-weight policies to balance the queues, by assigning greater service rates to links that have larger queue backlogs. For example, [21] contains a strong sample path optimality result for queue backlogs under stochastically symmetric traffic to parallel queues, and which is generalized in [6]. Additionally, [15] derives order optimal delay bounds when the arrival rates are 'f-balanced', and lie inside a scaled version of the stability region. Certain large deviations optimality results are also known [18], [19] for the class of max-weight policies when all the arrival processes are sufficiently well-behaved and light-tailed.

On the other hand, the traffic flows encountered in practice tend to be highly asymmetric, exhibiting wide range of variability or burstiness [11]. In this paper, we analyze the delay performance of generalized max-weight policies, when the competing traffic sources are highly asymmetric. We study a system consisting of two parallel queues, served by a single server. One of the queues is fed by a highly bursty arrival process, which is modeled as being heavy-tailed. The other queue is fed by a light-tailed arrival process. We refer to these queues as the 'heavy' and 'light' queues, respectively. The performance metric we focus on in the present paper is the tail behavior of the steady-state queue backlog distributions. This tail behavior essentially captures the probability of a large delay event occurring in

the queueing system.

To our knowledge, [12] was the first paper to study the performance of the max-weight family of scheduling policies, when heavy-tailed and light-tailed traffic compete for service. Specifically, it was shown in [12] that when the heavy-tailed traffic has an infinite variance, the light-tailed traffic experiences an infinite expected delay under max-weight scheduling, due to competition from the heavy-tailed traffic. The authors also studied a more general max-weight-α policy, wherein by increasing the preference afforded to the light queue, it is possible to make the expected delay of the light-tailed traffic finite. In a subsequent paper [9], we obtained an exact asymptotic characterization of the steady-state queue-backlog distributions under generalized max-weight policies, for a fairly general class of heavy-tailed distributions. Our results in [9] show that the light-tailed traffic always suffers a heavy-tailed backlog under max-weight-α scheduling, although the 'tail coefficient' can be influenced by appropriately adjusting the α parameters in the policy.

In [9], we assume that the queues are reliably connected to the server. In the present paper, we introduce channel variability into the model, and assume that the queues are connected to the server through time-varying ON/OFF links (Fig. 1). This can be viewed as a rudimentary model of a wireless uplink/downlink scenario, with two nodes communicating with a base station through fading channels. In this setup with time-varying ON/OFF links, we discover a *threshold phenomenon* with respect to the arrival rate of the light-tailed traffic, which essentially governs the queue backlog distribution faced by the light-tailed traffic. Under max-weight-α scheduling, we show that the light queue backlog distribution is light-tailed if the arrival rate to the light queue is below a certain threshold value, and heavy-tailed if the arrival rate is above the threshold value. This is in contrast with the case of reliable channels [9], where the light-tailed traffic faces heavy-tailed queue backlog for all positive arrival rates. Further, when the arrival rate is above the threshold value, we obtain the exact tail coefficient of the queue backlog distributions, which helps us identify all the bounded moments of the queue backlogs. This threshold behavior is intuitively due to the fact that the light-tailed traffic can always be served whenever the link serving the heavy-tailed queue is OFF, and is therefore guaranteed a minimum service rate independent of the behavior of the heavy queue.

The simplest way to guarantee a good tail behavior for the light queue distribution is to give the light queue complete priority over the heavy queue, so that it does not have to compete with the heavy queue for service. However, giving priority to the light queue has an important shortcoming – it is not a throughput optimal scheduling policy for the system. Indeed, we characterize the loss in throughput, and point out that giving complete priority to the light queue can cause undesirable instability effects
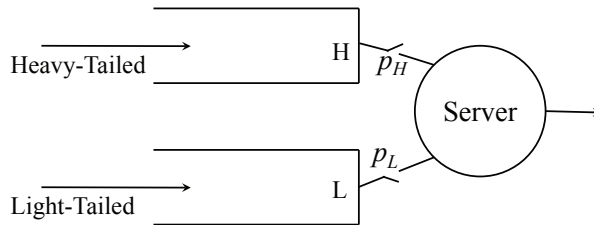
Fig. 1. A system of two parallel queues, with one of them fed with heavy-tailed traffic. The channels connecting the queues to the server are unreliable ON/OFF links.

in the heavy queue.

Thus on the one hand, the throughput optimal max-weight-α scheduling policy can lead to heavy-tailed asymptotics for the light queue. On the other hand, giving priority to the light queue leads to good tail behavior for the light queue, but is not throughput optimal. As a compromise, we study a log-max-weight (LMW) scheduling policy, which gives significantly more importance to the light queue compared to max-weight-α scheduling. We show that the LMW policy has both desirable attributes – namely, it is throughput optimal, and ensures good tail behavior for the light queue distribution.

The remainder of this paper is organized as follows. In Section II, we introduce the system model and the requisite technical preliminaries. In Section III, we study priority scheduling. Section IV deals with queue backlog behavior under max-weight-α scheduling. In Section V, we analyze the queue backlog behavior under log-max-weight scheduling. Section VI concludes the paper.

II. SYSTEM DESCRIPTION AND PRELIMINARIES

In this section, we describe the system model, and specify our assumptions about the traffic statistics. Our system consists of two parallel queues, H and L, served by a single server, as depicted in Fig. 1. Time is slotted, and stochastic arrivals of packet bursts occur to each queue in each slot. The server is capable of serving one packet per time slot, from only one of the queues according to a scheduling policy. Let $H(t)$ and $L(t)$ denote the number of packets that arrive at the end of slot t, to H and L respectively. Although we postpone the precise assumptions on the traffic statistics to Section II-B, let us loosely say that the input $L(t)$ is light-tailed, and $H(t)$ is heavy-tailed. We will refer to the queues H and L as the heavy and light queues, respectively.

The queues are connected to the server through time-varying links. Let $S_H(t) \in \{0, 1\}$ and $S_L(t) \in \{0, 1\}$ respectively denote the states of the channels connecting the H and L queues to the server. When a channel is in state 0, it is OFF, and no packets can be served from the corresponding queue in that

slot. When a channel is in state 1, it is ON, and a packet can be served from the corresponding queue if the server is assigned to that queue. This channel model can be used to represent fading wireless links in a two-user up-link or down-link system. We assume that the scheduler can observe the current channel states as well as the queue lengths before making a scheduling decision in a slot.

The channel processes $S_H(t)$ and $S_L(t)$ are independent of each other, and independent of the arrival processes. We assume that $S_H(t)$ and $S_L(t)$ are i.i.d. from slot to slot, distributed according to Bernoulli processes with positive means p_H and p_L respectively. That is, $\mathbb{P}\{S_i(\cdot) = 1\} = p_i, \ i \in \{H, L\}$. We say that a particular time slot t is *exclusive* to H, if $S_H(t) = 1$ and $S_L(t) = 0$, and similarly for L.

Before we specify the precise assumptions on the arrival processes, we pause to make some relevant definitions.

A. Heavy-tailed and light-tailed random variables

Definition 1: A non-negative random variable X is said to be *light-tailed* if there exists $\theta > 0$ for which $\mathbb{E}\left[\exp(\theta X)\right] < \infty$. A random variable is *heavy-tailed* if it is not light-tailed.

In other words, a light-tailed random variable is one that has a well defined moment generating function in a neighborhood of the origin. The complementary distribution function of a light-tailed random variable decays at least exponentially fast. Heavy-tailed random variables are those that have complementary distribution functions that decay slower than any exponential. We now define the tail coefficient of a random variable.

Definition 2: The *tail coefficient* of a random variable X is defined by

$$C_X = \sup\{c \geq 0 \mid \mathbb{E}\left[X^c\right] < \infty\}.$$

In words, the tail coefficient is the threshold where the power moment of a random variable starts to blow up. Note that the tail coefficient of a light-tailed random variable is infinite. On the other hand, the tail coefficient of a heavy-tailed random variable may be infinite (e.g., log-normal) or finite (e.g., Pareto). In this paper, we restrict our attention to the class of heavy-tailed random variables that have a finite tail coefficient.

We now state the precise assumptions on the arrival processes.

B. Assumptions on the arrival processes

1) The arrival processes $H(t)$ and $L(t)$ are independent of each other.
2) $H(t)$ is independent and identically distributed (i.i.d.) from slot-to-slot.

3) $L(t)$ is i.i.d. from slot-to-slot.

4) $L(\cdot)$ is light-tailed with $\mathbb{E}\left[L(t)\right] = \lambda_L$.

5) $H(\cdot)$ is heavy-tailed with tail coefficient C_H $(1 < C_H < \infty)$, and $\mathbb{E}\left[H(t)\right] = \lambda_H$.

The conditions for a rate pair (λ_H, λ_L) to be stably[1] supportable in this system are well known. Specifically, it follows from the results in [21] that the rate region of the system is given by

$$\Lambda = \{(\lambda_H, \lambda_L) \mid 0 \leq \lambda_L < p_L,\ 0 \leq \lambda_H < p_H,\ \lambda_H + \lambda_L < p_H + p_L - p_H p_L\}. \tag{1}$$

Thus, the rate region is pentagonal, with its boundary indicated by the solid line in Fig. 2.

Let $q_H(t)$ and $q_L(t)$, respectively, denote the number of packets in H and L at the beginning of slot t, under a particular scheduling policy, and let q_H and q_L denote the corresponding steady-state queue backlogs when they exist. Our aim is to characterize the distributions of q_H and q_L under various scheduling policies.

III. PRIORITY POLICIES

In this section, we study two extremal scheduling policies, namely, priority for L and priority for H. Our analysis helps us arrive at the conclusion that the tail of the heavy queue is inevitably heavy-tailed under any scheduling policy.

A. Priority for the heavy-tailed traffic

Under priority for H, the heavy queue receives service whenever it is non-empty and connected to the server. Queue L receives service during its exclusive slots, and when both queues are connected, but H is empty. It should be intuitively clear at the outset that this policy is bound to have an undesirable impact on the light queue. The reason we analyze this policy is that it gives us a best case scenario for the heavy queue. The following result shows that the heavy queue backlog distribution is one order heavier than its input distribution under this policy.

Proposition 1: Under priority for H, the steady-state queue backlog distribution of the heavy queue is a heavy-tailed random variable with tail coefficient equal to $C_H - 1$. That is, for every $\epsilon > 0$, we have

$$\mathbb{E}\left[q_H^{C_H-1-\epsilon}\right] < \infty, \tag{2}$$

and

$$\mathbb{E}\left[q_H^{C_H-1+\epsilon}\right] = \infty. \tag{3}$$

---

[1]The notion of stability we use is the positive recurrence of the system backlog Markov chain.
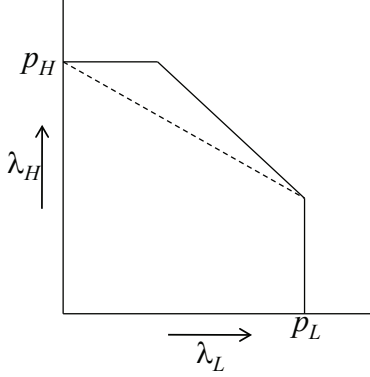
Fig. 2. The rate region of the system is shown in solid line, and the set of stabilizable rates under priority for L is the region under the dashed line.

Proof: We first note that under priority for H, the heavy queue behaves like a discrete time G/M/1 system. For such a queue, the upper bound (2) is easily obtained using a drift argument for the Lyapunov function $V(q_H(t)) = q_H(t)^{C_H - \epsilon}$. To obtain the lower bound (3), let us define $\tilde{H}$ as a fictitious heavy queue, which is fed by the same input sample path as the original queue H. However, $\tilde{H}$ is always connected to the server, and receives service in every slot. Notice now that q_H stochastically dominates $q_{\tilde{H}}$. It is therefore sufficient to show that $\mathbb{E}\left[q_{\tilde{H}}^{C_H - 1 + \epsilon}\right] = \infty$, and this can be accomplished by following [9, Theorem 1] . □

Since priority for H affords the most favorable treatment to the heavy queue, it follows that the tail behavior of H can be no better than the above under *any* policy.

Proposition 2: Under any scheduling policy, q_H is heavy-tailed with tail coefficient at most $C_H - 1$. That is, Equation (3) holds for *all* scheduling policies.

B. Priority for the light-tailed traffic

Under priority for L, the light queue is served whenever its channel is ON, and L is non-empty. The heavy queue is served during the exclusive slots of H, and in the slots when both channels are ON, but L is empty. This policy ensures that the light queue does not have to compete with the heavy queue for service, and guarantees the lowest possible light queue backlog among all policies. However, we show that this policy is not throughput optimal, and that it fails to stabilize the heavy queue for some arrival rates within the rate region in (1). The following theorem characterizes the behavior of both queues under priority for L.

Theorem 1: The following statements hold under priority scheduling for the light queue.

(i) If $\lambda_H > p_H(1 - \lambda_L)$, the heavy queue is *unstable*, and no steady-state exists.

(ii) If $\lambda_H < p_H(1 - \lambda_L)$, the heavy queue is stable, and its steady-state backlog q_H is heavy-tailed with tail coefficient $C_H - 1$.

(iii) q_L is light-tailed and satisfies the LDP

$$\lim_{b \to \infty} -\frac{1}{b} \log \mathbb{P}\{q_L > b\} = I_L,$$

where I_L is the *intrinsic exponent* of the light queue given by

$$I_L = \sup\left\{\theta \ \Big| \ \Lambda_L(\theta) - \log\left(1 - p_L + p_L e^{\theta}\right) < 0\right\}, \tag{4}$$

and $\Lambda_L(\theta) = \log \mathbb{E}\left[e^{\theta L(1)}\right]$ is the log moment generating function of $L(\cdot)$.

In Figure 2, the line $\lambda_H = p_H(1 - \lambda_L)$ is shown using a dashed segment. The above theorem asserts that H is stable under priority for L only in the trapezoidal region under the dashed line, while the rate region of the system is clearly larger. Therefore, priority for L is *not* throughput optimal in this setting. To summarize, priority for L can lead to instability of the heavy queue, but for all arrival rates that it *can* stabilize, the asymptotic behavior of both queues is as good as it can possibly be. Let us now prove the above theorem.

Proof: First, we note that the light queue behaves like a discrete time G/M/1 queue under priority, since the service time for each packet is geometrically distributed with mean $1/p_L$. Thus, q_L is light-tailed, and satisfies the same LDP as a G/M/1 queue. Statement (iii) therefore follows from classical large deviation results [5, Theorem 1.4].

Let us now prove statement (i) of the theorem. Under priority for L, denote by $\hat{D}_H(t) \in \{0, 1\}$ the indicator of a service opportunity afforded to the heavy queue in slot t. Thus, $\hat{D}_H(t) = 1$ if H is ON and the server is assigned to H during slot t, and zero otherwise. Note that $\hat{D}_H(t) = 1$ does not necessarily imply a departure from the heavy queue in that slot, since H could be empty. We will compute the long term average rate of service opportunities given to H under priority for L, defined as[2]

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \hat{D}_H(t).$$

Since the light queue behaves as a G/M/1 queue, the intervals between successive commencements of busy periods of L are renewal intervals. Let us denote by X_L a random variable representing the length of a renewal interval. Also denote by $\overline{B}$ and $\overline{I}$, respectively, the average length of a busy and idle period of L. The average length of a renewal interval is therefore $\mathbb{E}[X_L] = \overline{B} + \overline{I}$. Consider now

---

[2]We will see momentarily that this limit exists almost surely.

the total number of service opportunities $\hat{d}_H(i)$ given to H during the ith renewal interval. Thus, $\hat{d}_H(i)$ equals the number of exclusive slots of H during the renewal interval, plus the number of slots when both channels are ON and L is empty (the idle time). We can then consider $\hat{d}_H(i)$ as a renewal reward function, and invoke the renewal reward theorem [4] to write (almost surely)

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \hat{D}_H(t) = \frac{\mathbb{E}\left[\hat{d}_H(i)\right]}{\overline{B} + \overline{I}}.$$

Let us now compute $\mathbb{E}\left[\hat{d}_H(i)\right]$. First, the average number of exclusive slots of H during a renewal interval is given by $p_H(1 - p_L)(\overline{B} + \overline{I})$. Second, the average number of slots when both channels are ON, and L is empty is given by $\overline{I} p_H p_L$. Therefore, $\mathbb{E}\left[\hat{d}_H(i)\right] = p_H(1 - p_L)(\overline{B} + \overline{I}) + \overline{I} p_H p_L$. Substituting this in the reward theorem, we get

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \hat{D}_H(t) = p_H(1 - p_L) + p_H p_L \frac{\overline{I}}{\overline{B} + \overline{I}}. \tag{5}$$

In the above, note that $\frac{\overline{I}}{\overline{B} + \overline{I}}$ is the fraction of time that the light queue is idle. Also, by applying Little's law to the server at the light queue, we find that the fraction of time the light queue is busy should equal $\frac{\lambda_L}{p_L}$, which is the load on the light queue. Therefore,

$$\frac{\overline{I}}{\overline{B} + \overline{I}} = 1 - \frac{\lambda_L}{p_L}.$$

Substituting in (5),

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \hat{D}_H(t) = p_H(1 - p_L) + p_H p_L \left(1 - \frac{\lambda_L}{p_L}\right) = p_H(1 - \lambda_L). \tag{6}$$

Thus, the average rate of service opportunities for H almost surely equals $p_H(1 - \lambda_L)$. If $\lambda_H > p_H(1 - \lambda_L)$, then the average service rate given to the heavy queue is dominated by the average arrival rate, leading to the instability of H. This proves statement (i).

Finally, to prove statement (ii), we can use a direct Lyapunov approach as shown in [7, Theorem 5.1]. Alternatively, we could invoke [3, Theorem 1] on the heavy queue in isolation, after verifying that the requisite conditions are met. □

The special case in which the queues are always connected to the server, i.e., $p_H = p_L = 1$, is interesting. In this case, the set of arrival rates stabilizable under priority for L coincides with the stability region of the system, which is given by

$$\{(\lambda_H, \lambda_L) \mid \lambda_H + \lambda_L < 1\}.$$

Therefore, when the queues are reliably connected to the server, priority scheduling for the light-tailed traffic is throughput optimal, and also ensures the best possible tail behavior for both queues.

IV. MAX-WEIGHT-α SCHEDULING

In this section, we analyze the tail behavior of the light queue distribution under max-weight-α scheduling. For fixed parameters $\alpha_H > 0$ and $\alpha_L > 0$, the max-weight-α policy operates as follows. During each slot t, compare

$$q_L(t)^{\alpha_L} S_L(t) \gtreqless q_H(t)^{\alpha_H} S_H(t),$$

and serve one packet from a queue that wins the comparison. Ties are broken in favor of the light queue. Note that $\alpha_L = \alpha_H$ corresponds to the usual max-weight policy, which serves the longest connected queue in each slot. The case $\alpha_L/\alpha_H > 1$ corresponds to emphasizing the light queue over the heavy queue, and vice-versa.

It can be shown using standard Lyapunov arguments that max-weight-α scheduling is throughput optimal for all $\alpha_H > 0$ and $\alpha_L > 0$. That is, it can stably support all arrival rates within the rate region (1). This throughput optimality result follows, for example, from [3, Theorem 1].

We show that under max-weight-α scheduling, the tail behavior of the steady-state light queue backlog distribution is strongly dependent on λ_L, the arrival rate to the light queue. Specifically, we show that q_L is light-tailed when λ_L is below a threshold value, and heavy-tailed with a finite tail coefficient for λ_L above the threshold value.

The following result shows that the light queue distribution is light-tailed under any 'reasonable' policy, as long as the rate λ_L is smaller than a threshold value.

Proposition 3: Suppose that $\lambda_L < p_L(1 - p_H)$. Then q_L is light-tailed under *any* policy that serves L during its exclusive slots.

Proof: The proof is straightforward once we note that the exclusive slots of L occur independently during each slot with probability $p_L(1 - p_H)$. Indeed, consider the L queue under a policy that serves L *only* during its exclusive slots. Under this policy, the L queue behaves like a G/M/1 queue with light-tailed inputs at rate λ_L, and service rate $p_L(1 - p_H)$. It can be shown using standard large deviation arguments [5, Theorem 1.4] that q_L is light-tailed under the policy that serves L only during its exclusive slots. It follows, using a stochastic dominance argument that q_L is light-tailed under any policy that serves L during its exclusive slots. $\qquad\square$

The above proposition implies that for $\lambda_L < p_L(1 - p_H)$, the light queue distribution is light-tailed under max-weight-α scheduling. The region $\lambda_L < p_L(1 - p_H)$ is shown unshaded in Fig. 3. Thus, q_L is light-tailed under max-weight-α scheduling for arrival rates in the unshaded region.
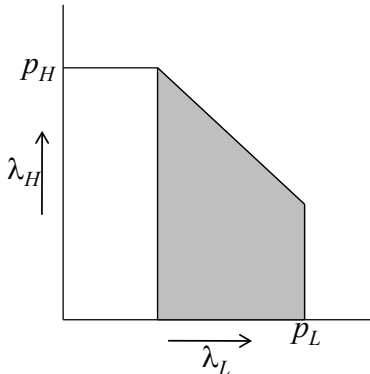
Fig. 3. Under max-weight-α scheduling, q_L is light-tailed for arrival rates in the unshaded region, and heavy-tailed in the shaded region.

In the remainder of this section, we investigate the tail behavior of the light queue under max-weight-α scheduling when the arrival rate is above the threshold, i.e., for $\lambda_L > p_L(1 - p_H)$. In this case, the light queue receives traffic at a higher rate than can be supported by the exclusive slots of L alone. Therefore, the light queue has to compete for service with the heavy queue during the slots that both channels are ON. Since the heavy queue can be very large with substantially high probability, it seems intuitively reasonable that the light queue will suffer from this competition, and also take on a heavy-tailed behavior. This intuition is indeed correct, although proving the result takes some effort.

We prove that the light queue distribution is heavy-tailed when $\lambda_L > p_L(1 - p_H)$ for all values of the scheduling parameters α_L and α_H. We also obtain the exact tail coefficient of the light queue distribution for 'plain' max-weight scheduling ($\alpha_L/\alpha_H = 1$), and for the regime where the light queue is given more importance ($\alpha_L/\alpha_H > 1$).

A. Max-weight scheduling

Let us first characterize the tail coefficient of the steady-state light queue backlog under the max-weight policy, which serves the longest connected queue in each slot. Since q_L is light-tailed for $\lambda_L < p_L(1 - p_H)$ according to Proposition 3, we will focus on the case $\lambda_L > p_L(1 - p_H)$.

Theorem 2: Suppose that $\lambda_L > p_L(1 - p_H)$. Then, under max-weight scheduling, q_L is heavy-tailed with tail coefficient $C_H - 1$.

In terms of Fig. 3, the theorem asserts that q_L is heavy-tailed with tail coefficient $C_H - 1$ for all arrival rates in the shaded region. Proving the above result involves showing (i) an upper bound: $\mathbb{E}\left[q_H^{C_H - 1 - \epsilon}\right] < \infty$, and (ii) a lower bound: $\mathbb{E}\left[q_H^{C_H - 1 + \epsilon}\right] = \infty$, for any $\epsilon > 0$. We deal with each part

separately.

1) Upper Bound for max-weight scheduling:

Proposition 4: Under max-weight scheduling, we have

$$\mathbb{E}\left[q_L^{C_H-1-\epsilon}\right] < \infty, \ \forall \ \epsilon > 0.$$

Proof: This is a special case ($\alpha_L/\alpha_H = 1$) of Proposition 6, in the next section. □

2) Lower Bound for max-weight scheduling:

Proposition 5: Suppose that $\lambda_L > p_L(1 - p_H)$. Then, under max-weight scheduling, we have

$$\mathbb{E}\left[q_L^{C_H-1+\epsilon}\right] = \infty, \ \forall \ \epsilon > 0.$$

This is a special case of Proposition 7, but we will provide a proof because this special case is more transparent. Since the proof is rather involved, we describe the idea informally, and present the formal proof in Appendix A. In our intuitive argument, we will argue that

$$\lim_{t\to\infty} \mathbb{E}\left[q_L(t)^{C_H-1+\epsilon}\right] = \infty. \tag{7}$$

The above is the limit of the expectation of a sequence of random variables, whereas what we really want in Proposition 5 is the expectation of the limiting random variable q_L. Although it is by no means obvious that the limit and the expectation can be interchanged here, we will ignore this as a technical point for the time being.

The main idea behind the proof is to consider the renewal intervals that commence at the beginning of each busy period of the system. Let us define the renewal reward process $R(t) = q_L(t)^{C_H-1+\epsilon}$. By the key renewal theorem for arithmetic processes [4, pp. 81],

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

where $\mathbb{E}\left[R\right]$ denotes the expected reward accumulated over a renewal interval, and $\mathbb{E}\left[T\right] < \infty$ is the mean renewal interval. It is therefore enough to show that[3]

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H-1+\epsilon}\right] = \infty.$$

To see intuitively why the above expectation is infinite, let us condition on the busy period commencing at time 0 with a burst of size b to the heavy queue[4]. After this instant, the heavy queue drains at rate p_H, assuming for the sake of a lower bound that there are no further bursts arriving at H. In the

---

[3]Without loss of generality, we have considered a busy period that commences at time 0.

[4]It is easy to show that this event has positive probability for all large enough b.

meantime, the light queue receives traffic at rate λ_L, and gets served only during the exclusive slots of L, which occur at rate $p_L(1 - p_H)$. With high probability therefore, the light queue will steadily build up at rate $\lambda_L - p_L(1 - p_H)$, until it eventually catches up with the draining heavy queue. It can be shown that the light queue will build up to an $\Omega(b)$ level before it catches up with the heavy queue. Further, the light queue backlog stays at $\Omega(b)$ for a time interval of length $\Omega(b)$. Therefore, with high probability, the reward is at least $\Omega(b^{C_H - 1 + \epsilon})$ for $\Omega(b)$ time slots. Thus, for some constant K,

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H - 1 + \epsilon}\right] \geq \mathbb{E}\left[Kb \cdot b^{C_H - 1 + \epsilon}\right] = \mathbb{E}\left[Kb^{C_H + \epsilon}\right] = \infty,$$

where the last expectation is infinite because the initial burst size has tail coefficient equal to C_H.

In words, the light queue not only grows to a level proportionate to the initial burst size, but also stays large for a period of time that is proportional to the burst size. This leads to a light queue distribution that is one order heavier than the burst size distribution.

B. Max-weight-α scheduling with $\alpha_L \geq \alpha_H$

In this subsection, we characterize the exact tail coefficient of the light queue distribution under max-weight-α scheduling, with $\alpha_L \geq \alpha_H$. We only treat the case $\lambda_L > p_L(1 - p_H)$, since q_L is known to be light-tailed otherwise. Our main result for this regime is the following.

Theorem 3: Suppose that $\lambda_L > p_L(1 - p_H)$. Then, under max-weight-α scheduling with $\alpha_L \geq \alpha_H$, q_L is heavy-tailed with tail coefficient

$$\gamma = \frac{\alpha_L}{\alpha_H}(C_H - 1). \tag{8}$$

In terms of Fig. 3, the above theorem asserts that q_L is heavy-tailed with tail coefficient γ for all arrival rates in the shaded region. As before, proving this result involves showing (i) an upper bound of the form $\mathbb{E}\left[q_H^{\gamma - \epsilon}\right] < \infty$, and (ii) a lower bound of the form $\mathbb{E}\left[q_H^{\gamma + \epsilon}\right] = \infty$, for all $\epsilon > 0$. We deal with each of them separately.

1) Upper Bound for max-weight-α scheduling:

Proposition 6: Under max-weight-α scheduling, we have

$$\mathbb{E}\left[q_L^{\gamma - \epsilon}\right] < \infty, \ \forall \ \epsilon > 0.$$

Proof: The result is a consequence of a theorem in [3]. Indeed, max-weight-α scheduling in our context is equivalent to comparing $q_L(t)^{\beta \alpha_L} S_L(t)$ with $q_H(t)^{\beta \alpha_H} S_H(t)$, where $\beta > 0$ is arbitrary, and scheduling

the winning queue in each slot. In particular, if we choose $\beta = (C_H - 1)/\alpha_H - \epsilon/\alpha_L$, the conditions imposed in [3, Theorem 1] are satisfied for any $\epsilon > 0$, so that the steady-state queue backlogs satisfy

$$\mathbb{E}\left[q_L^{\gamma - \epsilon}\right] < \infty,$$

and

$$\mathbb{E}\left[q_H^{C_H - 1 - \frac{\alpha_H}{\alpha_L}\epsilon}\right] < \infty. \tag{9}$$

$\square$

Remark 1:   (i) Proposition 6 is valid for any parameters α_L and α_H, and not just for $\alpha_L \geq \alpha_H$.

(ii) Equation (9) and Proposition 2 together imply that the tail coefficient of q_H is equal to $C_H - 1$ under max-weight-α scheduling, for any parameters α_L and α_H.

2) Lower Bound for max-weight-α scheduling with $\alpha_L \geq \alpha_H$:

Proposition 7: Suppose that $\lambda_L > p_L(1 - p_H)$. Then, under max-weight-α scheduling with $\alpha_L \geq \alpha_H$, we have

$$\mathbb{E}\left[q_L^{\gamma + \epsilon}\right] = \infty, \ \forall \ \epsilon > 0.$$

To prove the above result, we take an approach that is conceptually similar to the proof of Proposition 5. We consider the renewal process that commences at the beginning of each busy period of the system, and define the reward process $R_\gamma(t) = q_L(t)^{\gamma + \epsilon}$. We will show that the expected reward accumulated over a renewal interval is infinite. The key renewal theorem will then imply that $\lim_{t \to \infty} \mathbb{E}\left[q_L(t)^{\gamma + \epsilon}\right] = \infty$. Finally, the result we want can be obtained by invoking a truncation argument to interchange the limit and the expectation.

To see intuitively why the expected reward over a renewal interval is infinite, let us condition on the busy period commencing with a burst of size b at the heavy queue. Starting at this instant, the light queue will build up at the rate $\lambda_L - p_L(1 - p_H)$ with high probability. However, unlike in the case of max-weight scheduling, the light queue only builds up to an $\Omega(b^{\alpha_H/\alpha_L})$ level before it 'catches up' with the heavy queue and wins back the service preference. It can also be shown that the light queue catches up within a time interval of length $\Omega(b^{\alpha_H/\alpha_L})$. It might therefore be tempting to argue that the light queue stays above $\Omega(b^{\alpha_H/\alpha_L})$ for an interval of duration $\Omega(b^{\alpha_H/\alpha_L})$. Although this argument is not incorrect as such, it fails to capture what typically happens in the system. Let us briefly follow through with this argument, in order to understand why it does not give us the lower bound we want.

Indeed, following the above argument, the reward is at least $\Omega(b^{(\gamma+\epsilon)\alpha_H/\alpha_L}) = \Omega(b^{C_H-1+\epsilon\alpha_H/\alpha_L})$ for $\Omega(b^{\alpha_H/\alpha_L})$ time slots, so that the expected reward over the renewal interval is lower bounded by

$$\mathbb{E}_b\left[\Omega(b^{\alpha_H/\alpha_L})\Omega(b^{C_H-1+\epsilon\alpha_H/\alpha_L})\right] = \mathbb{E}_b\left[\Omega(b^{C_H-1+\alpha_H/\alpha_L+\epsilon\alpha_H/\alpha_L})\right].$$

However, the right-hand side above turns out to be *finite* for $\alpha_L/\alpha_H > 1$. Therefore, the above simple bound fails to give the result we are after.

The problem with the above argument is that it looks at the time scale at which the light queue *catches up*, whereas the event that decides the tail coefficient happens *after* the light queue catches up. In particular, the light queue catches up relatively quickly, in a time scale of $\Theta(b^{\alpha_H/\alpha_L})$. However, after the light queue catches up with the heavy queue, the two queues drain together, with most of the slots being used to serve the heavy queue. In fact, as we show, before the light queue backlog can drain by a constant factor after catch-up, the heavy queue drains by $\Omega(b)$. As such, the light queue remains at an $\Omega(b^{\alpha_H/\alpha_L})$ level for $\Omega(b)$ time slots. Therefore, the expected reward can be lower bounded by

$$\mathbb{E}_b\left[\Omega(b)b^{C_H-1+\epsilon\alpha_H/\alpha_L}\right] = \mathbb{E}_b\left[\Omega(b^{C_H+\epsilon\alpha_H/\alpha_L})\right] = \infty,$$

which is what we want. In sum, the light queue builds up relatively quickly until catch-up, but takes a long time to drain out after catch-up. The proof is relegated to Appendix B.

C. Max-weight-α scheduling with $\alpha_L < \alpha_H$

We finally consider the case $\alpha_L < \alpha_H$ under max-weight-α scheduling, and study the asymptotic behavior of q_L. Recall that max-weight-α scheduling with $\alpha_L < \alpha_H$ corresponds to giving the heavy queue more importance compared to the light queue. In this regime, we show that q_L is heavy-tailed with a finite tail coefficient, for arrival rates in the shaded region of Figure 3.

Our first result for this case is an upper bound on the tail coefficient of q_L. Intuitively, we would expect that the tail behavior of q_L in this regime cannot be better than it is under max-weight scheduling. In other words, the tail coefficient of q_L in this regime cannot be larger than $C_H - 1$. This intuition is indeed correct.

Proposition 8: Suppose that $\lambda_L > p_L(1-p_H)$. Then, under max-weight-α scheduling with $\alpha_L < \alpha_H$, the tail coefficient of q_L is at most $C_H - 1$.

Proof: The argument is similar to the proof of Proposition 5. Specifically, conditioning on an initial burst of size b arriving to the heavy queue, it can be shown that with high probability, q_L will be $O(b)$ in size for at least $O(b)$ time slots. □
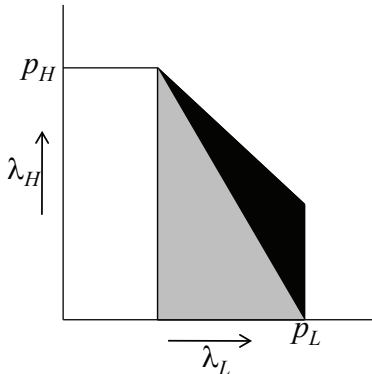
Fig. 4. Under max-weight-α scheduling with $\alpha_L < \alpha_H$, q_L is light-tailed for arrival rates in the unshaded region, and heavy-tailed with tail coefficient equal to $C_H - 1$ in for arrival rates in the gray region. For arrival rates in the region colored black, the tail coefficient lies in $[\gamma, C_H - 1]$.

Next, to obtain a lower bound on the tail coefficient of q_L, recall that Proposition 6 holds for the present regime as well. Thus,

$$\gamma = \frac{\alpha_L}{\alpha_H}(C_H - 1)$$

is a lower bound[5] on the tail coefficient of q_L. In sum, we have shown that for $\lambda_L > p_L(1 - p_H)$, the light queue backlog distribution is heavy-tailed, with a tail coefficient that lies in the interval $[\gamma, C_H - 1]$.

It turns out that we can obtain the exact tail coefficient of q_L for arrival rates in a *subset* of the shaded region in Fig. 3. Specifically, consider the region represented by $p_L(1 - p_H) < \lambda_L < p_L(1 - \lambda_H)$. In Fig. 4, this region is shown in gray. It can be shown that all arrival rates in the region shaded gray can be stabilized under priority for H. Furthermore, under priority for H, it can be shown that q_L is heavy-tailed with tail coefficient equal to $C_H - 1$, when $p_L(1 - p_H) < \lambda_L < p_L(1 - \lambda_H)$.

Since the tail of q_L under max-weight-α scheduling with any parameters is no worse than under priority for H, we can conclude that the tail coefficient of q_L is at least $C_H - 1$ when $p_L(1 - p_H) < \lambda_L < p_L(1 - \lambda_H)$. Combining this with Proposition 8, we conclude that the tail coefficient q_L is *equal* to $C_H - 1$, when the arrival rate pair lies in the gray region of Fig. 4.

Proposition 9: Suppose that $p_L(1 - p_H) < \lambda_L < p_L(1 - \lambda_H)$. Then, under max-weight-α scheduling with $\alpha_L < \alpha_H$, the tail coefficient of q_L is equal to $C_H - 1$.

The region shaded black in Fig. 4 ($\lambda_L > p_L(1 - \lambda_H)$) corresponds to the arrival rates for which priority for H is *not* stabilizing[6]. Under max-weight-α scheduling with $\alpha_L < \alpha_H$, we are unable to

---

[5]Note that γ is *smaller* than $C_H - 1$ in this regime.

[6]This case is symmetric to the case in Theorem 1(i).

determine the exact tail coefficient of q_L for arrival rates in the black region of Fig. 4. However, we have shown that the tail coefficient lies in the interval $[\gamma, C_H - 1]$.

D. Special case of reliable links

In the special case of reliably connected links ($p_H = p_L = 1$), the tail behavior of q_L under max-weight-α scheduling can be obtained from our foregoing analysis. Specifically, it follows from the results above that the light queue backlog distribution is heavy-tailed under max-weight-α scheduling, for all values of the scheduling parameters, and all non-zero arrival rates. The tail coefficient of q_L for this case is given by

(i) $C_H - 1$ for $\frac{\alpha_L}{\alpha_H} \leq 1$, and

(ii) $\gamma = \frac{\alpha_L}{\alpha_H}(C_H - 1)$ for $\frac{\alpha_L}{\alpha_H} > 1$.

We remark that this recovers our results in [9].

E. Section summary

We showed the following threshold result in this section. When $\lambda_L < p_L(1 - p_H)$, the light queue backlog distribution is light-tailed under max-weight-α scheduling, for all values of the scheduling parameters. However, when $\lambda_L > p_L(1 - p_H)$, the light queue distribution is inevitably heavy-tailed under max-weight-α scheduling. In particular, under max-weight scheduling ($\alpha_L = \alpha_H$), the tail coefficient of q_L is equal to $C_H - 1$. For $\alpha_L \geq \alpha_H$, the tail coefficient of q_L is $\gamma = (C_H - 1)\alpha_L/\alpha_H$. Finally, for $\alpha_L < \alpha_H$, the tail coefficient of q_L lies in $[\gamma, C_H - 1]$. Finally, we also showed that the heavy queue distribution is heavy-tailed with tail coefficient $C_H - 1$ for all values of the scheduling parameters.

V. Log-Max-Weight Scheduling

In this section, we study the performance of log-max-weight scheduling policy. During each time slot t, the log-max-weight policy compares

$$q_L(t)S_L(t) \gtreqless \log(1 + q_H(t))S_H(t),$$

and serves one packet from the queue that wins the comparison. Again, ties are broken to favor the light queue.

The main idea in the LMW policy is to give preference to the light queue to a far greater extent than any max-weight-α policy. Specifically, for $\alpha_L/\alpha_H > 1$, the max-weight-α policy compares q_L to a

power of q_H that is smaller than 1. On the other hand, LMW scheduling compares q_L to a logarithmic function of q_H, leading to a significant preference for the light queue. We will show that this significant de-emphasis of the heavy queue with respect to the light queue ensures a better tail behavior for q_L compared to max-weight-α scheduling.

Furthermore, the LMW policy has another useful property when the heavy queue gets overwhelmingly large. Although the LMW policy significantly de-emphasizes the heavy queue, it does not *ignore* it, unlike priority for L. That is, if the H queue backlog gets overwhelmingly large compared to L, the LMW policy will serve the heavy queue. In contrast, priority for L will ignore any build-up in H, as long as L is non-empty. This property ensures that the LMW policy stabilizes all arrival rates within the rate region in (1).

We show that LMW scheduling has desirable performance on both fronts, namely throughput optimality, and the tail behavior of the light queue backlog. The LMW policy can be shown to be throughput optimal, using the results in [3]. In terms of the tail, we show that the LMW policy guarantees that the light queue backlog distribution is light-tailed, for all arrival rates that can be stabilized by priority for L. For arrival rates that are not stabilizable under priority for L, the LMW policy will still stabilize the system, although we are not able to guarantee that q_L is light-tailed for these arrival rates.

Let us now state the main result regarding LMW scheduling.

Theorem 4: Under LMW scheduling, q_L is light-tailed if at least one of the following conditions hold:

 (i) $\lambda_L < p_L(1 - p_H)$, or

(ii) $\lambda_H < p_H(1 - \lambda_L)$.

Note that for $\lambda_L < p_L(1 - p_H)$, q_L is easily seen to be light-tailed under LMW scheduling, since the arrival rate is small enough to be supported by the exclusive slots of L. The second condition in Theorem 4 states that for all arrival rates that can be stabilized under priority for L (i.e., the trapezoidal region in Fig. 2), q_L is light-tailed under LMW scheduling.

The union of the two regions in which q_L is light-tailed according to Theorem 4 is shown unshaded in Fig. 5. As can be seen, the unshaded region occupies most of the rate region, except for the shaded triangle. For arrival rates in the shaded triangle, the LMW policy still stabilizes the system. However, we are unable to determine the tail behavior of q_L for arrival rates in the shaded triangle.

The proof of Theorem 4 is omitted in the interest of brevity, and can be found in [7, Theorem 5.4]. We remark that in a recent paper [14], it has been proven that q_L is light-tailed under the LMW policy for *all* arrival rates inside the rate region. This more general proof uses a completely different approach,
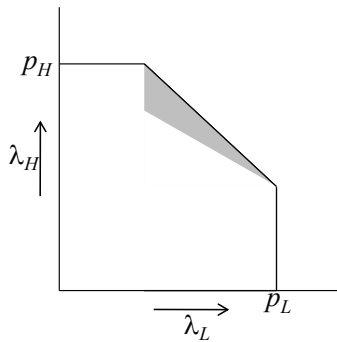
Fig. 5.   Under LMW scheduling, q_L is light-tailed for arrival rates in the unshaded region.

and can be found in [14, Theorem 9].

VI. CONCLUSIONS

We considered a system of parallel queues fed by a mix of heavy-tailed and light-tailed traffic, and served by a single server through time-varying channels. We studied the tail behavior of the queue backlog distributions under various scheduling policies. We showed that the backlog distribution of the heavy queue is inevitably heavy-tailed. In contrast, the light queue backlog distribution can be heavy-tailed or light-tailed, depending on the arrival rates and the scheduling policy. A major contribution of this paper is the characterization of the tail of the queue backlog distributions under max-weight-α scheduling. We showed that the light queue backlog distribution under max-weight-α scheduling is light-tailed for arrivals rates below a certain threshold, and heavy-tailed for arrival rates above the threshold.

Another contribution of the paper was to show that the LMW scheduling policy ensures that the light queue backlog distribution is light-tailed, in addition to being throughput optimal. Indeed, we believe that the LMW policy occupies a special place in the context of scheduling light-tailed traffic in the presence of heavy-tailed traffic. This is because the LMW policy de-emphasizes the heavy-tailed flow sufficiently to maintain good light queue asymptotics, while also ensuring network-wide stability.

APPENDIX A

PROOF OF PROPOSITION 5

We will first show Eq. (7) and then use a truncation argument to interchange the limit and the expectation. Consider the renewal process defined by the commencement of each busy period of the

system. Let T denote a typical renewal interval. We have $\mathbb{E}[T] < \infty$ since the system is stable. Define the reward function

$$R(t) = q_L(t)^{C_H - 1 + \epsilon}.$$

As argued earlier, due the key renewal theorem, it is enough to show that the expected reward accumulated over a renewal interval is infinite. Without loss of generality, let us consider a busy period that commences at time 0. We need to show that

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H - 1 + \epsilon}\right] = \infty.$$

The busy period that commences at time 0 can be of three different types. It can commence with (i) a burst arriving to L alone, or (ii) a burst arriving to H alone, or (iii) bursts arriving to both H and L simultaneously. It can be shown that all the three events have positive probabilities[7]. The event that is of interest to us is (ii), i.e., the busy period commencing with a burst at the heavy queue only, so that $q_H(0) > 0$ and $q_L(0) = 0$. Let us denote this event by $\mathcal{E}_H = \{q_H(0) > 0, q_L(0) = 0\}$. We now have the following lower bound

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H - 1 + \epsilon}\right] \geq \mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H - 1 + \epsilon}; \ \mathcal{E}_H\right] = \mathbb{E}_b\left[\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H - 1 + \epsilon}; \ \mathcal{E}_H \Big| q_H(0) = b\right]\right].$$

In the last step above, we have iterated the expectation over the initial burst size b. The inner expectation above is a function of b; let us denote it by

$$g_\epsilon(b) := \mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H - 1 + \epsilon}; \ \mathcal{E}_H \Big| q_H(0) = b\right].$$

Thus,

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H - 1 + \epsilon}\right] \geq \mathbb{E}_b[g_\epsilon(b)] \geq \mathbb{E}_b[g_\epsilon(b); \ b > b_0], \ \forall b_0 \geq 1. \tag{10}$$

Since the above bound is true for any b_0, we can make b_0 as large as we want. In particular, we will make the initial burst size large enough to be able to assert that the arrival process to L as well as the channel processes behave 'typically' for time scales of order b.

To be more precise, choose $\delta > 0$ such that $\lambda_L - p_L(1 - p_H) - 3\delta = \eta > 0$, and choose any small $\kappa > 0$. Define

$$\tau_b = \frac{b}{2(p_H + \lambda_L)}.$$

---

[7]In fact, we can explicitly compute the probability of each of the three events in terms of the probability mass at 0 for $H(\cdot)$ and $L(\cdot)$, but the actual probabilities are not important for the proof.

For large enough b_0, and $b > b_0$, it is clear from the (weak) law of large numbers (LLN) that

$$\mathbb{P}\left\{\left|\frac{1}{\tau_b}\sum_{i=0}^{\tau_b} S_H(i) - p_H\right| > \delta\right\} < \kappa.$$

In words, the channel process of H is overwhelmingly likely to behave according to its mean p_H. Now for all $t \leq \tau_b$, the backlog of H can be lower bounded as

$$q_H(t) \geq b - \sum_{i=0}^{\tau_b} S_H(i) \geq b - (p_H + \delta)\tau_b = b\left(\frac{p_H + 2\lambda_L - \delta}{2(p_H + \lambda_L)}\right), \tag{11}$$

with probability greater than $1 - \kappa$. Similarly, the input process to the light queue is also likely to behave according to its mean. That is, for large enough b_0 and $b > b_0$,

$$\mathbb{P}\left\{\left|\frac{1}{\tau_b}\sum_{i=0}^{\tau_b} L(i) - \lambda_L\right| > \delta\right\} < \kappa.$$

Therefore, for all $t \leq \tau_b$, the backlog of L can be upper bounded as

$$q_L(t) \leq \sum_{i=0}^{\tau_b} L(i) \leq b\left(\frac{\lambda_L + \delta}{2(p_H + \lambda_L)}\right), \tag{12}$$

with probability greater than $1 - \kappa$. From (11), (12), and the independence of the processes $L(\cdot)$ and $S_H(\cdot)$, we can conclude that $q_H(t) > q_L(t)$ for all $t \leq \tau_b$, with probability greater than $1 - 2\kappa$. Since the light queue remains smaller that the heavy queue for $t \leq \tau_b$ with high probability, it follows that the light queue receives service only during its exclusive slots. More precisely, the departure process from the light queue can be bounded as

$$\sum_{i=1}^{\tau_b} D_L(i) \leq \sum_{i=1}^{\tau_b} S_L(i)(1 - S_H(i)),$$

with probability at least $1 - 2\kappa$. However, the exclusive slots of L are also overwhelmingly likely to behave according to the mean:

$$\mathbb{P}\left\{\left|\frac{1}{\tau_b}\sum_{i=0}^{\tau_b} S_L(i)(1 - S_H(i)) - p_L(1 - p_H)\right| > 2\delta\right\} < \kappa.$$

Thus,

$$\sum_{i=1}^{\tau_b} D_L(i) \leq \tau_b(p_L(1 - p_H) + 2\delta),$$

with probability at least $1 - 3\kappa$. Using the above bound on the departures from L, along with the fact that arrivals to L are also typical, we can lower bound $q_L(\tau_b)$ with high probability. Indeed,

$$q_L(\tau_b) = \sum_{i=1}^{\tau_b} L(i) - \sum_{i=1}^{\tau_b} D_L(i) \geq \tau_b(\lambda_L - \delta) - \tau_b(p_L(1 - p_H) + 2\delta) = b\left(\frac{\eta}{2(p_H + \lambda_L)}\right),$$

with probability at least $1 - 3\kappa$. Next, since at most one packet can be served in a slot, we have,

$$q_L(t) \geq b\left(\frac{\eta}{4(p_H + \lambda_L)}\right), \quad \text{if} \quad \tau_b \leq t \leq \tau_b + \tau_b\left(\frac{\eta}{2}\right),$$

with probability at least $1 - 3\kappa$.

We can thus lower bound $g_\epsilon(b)$ for large enough b_0 and $b > b_0$ as

$$g_\epsilon(b)1_{\{b>b_0\}} = \mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H-1+\epsilon}; \, \mathcal{E}_H \Big| q_H(0) = b\right] 1_{\{b>b_0\}}$$

$$\geq \left[(1-3\kappa)\sum_{i=\tau_b}^{\tau_b+\tau_b\left(\frac{\eta}{2}\right)} \left(\frac{\eta b}{4(p_H + \lambda_L)}\right)^{C_H-1+\epsilon}\right] 1_{\{b>b_0\}}$$

$$\geq (1-3\kappa)\frac{\eta\tau_b}{2}\left(\frac{\eta\tau_b}{2}\right)^{C_H-1+\epsilon} 1_{\{b>b_0\}} = Kb^{C_H+\epsilon}1_{\{b>b_0\}}, \tag{13}$$

for some constant $K > 0$. Thus, going back to (10),

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{C_H-1+\epsilon}\right] \geq \mathbb{E}_b\left[g_\epsilon(b); \, b > b_0\right] \geq \mathbb{E}_b\left[Kb^{C_H+\epsilon}; \, b > b_0\right] = \infty.$$

The last step is because the initial burst size b has tail coefficient C_H, so that $\mathbb{E}_b\left[b^{C_H+\epsilon}; \, b > b_0\right] \geq \mathbb{E}_b\left[b^{C_H+\epsilon} - b_0^{C_H+\epsilon}\right] = \infty$ for all b_0. Therefore, we are done proving (7).

Finally, we use a truncation argument to prove that $\mathbb{E}\left[q_L^{C_H-1+\epsilon}\right] = \infty$, where q_L is the steady-state limit of $q_L(t)$.

A. Truncation argument

Our intention is to show that the limit and the expectation in (7) can be interchanged, so that we get the desired moment result for the limiting random variable q_L. Our truncation argument relies on the Monotone Convergence Theorem (MCT) [1, Theorem 16.2], as well as a result that affirms the convergence of moments when there is convergence in distribution [1, Theorem 25.12].

The main idea here is to define a truncated reward function

$$R_M(t) = (M \wedge q_L(t))^{C_H-1+\epsilon},$$

where M is a large integer, and $M \wedge q_L(t) := \min(M, q_L(t))$. There are three steps in our truncation argument.

(i) Tracing all the steps leading up to (13) in the proof above, and using the key renewal theorem for the truncated reward function, we can show that

$$w_M := \lim_{t \to \infty} \mathbb{E}\left[R_M(t)\right] \geq \frac{1 - 3\kappa}{\mathbb{E}\left[T\right]} \mathbb{E}_b\left[\frac{\eta \tau_b}{2}\left(M \wedge \left(\frac{\eta \tau_b}{2}\right)\right)^{C_H - 1 + \epsilon} 1_{\{b > b_0\}}\right], \tag{14}$$

for all M and large enough b_0. The left hand side in the above inequality is a function of M, which we have denoted by w_M. The expression inside the expectation on the right is a function of b and M, which we denote by

$$u_M(b) = \frac{\eta \tau_b}{2}\left(M \wedge \left(\frac{\eta \tau_b}{2}\right)\right)^{C_H - 1 + \epsilon} 1_{\{b > b_0\}}.$$

When viewed as a sequence of functions indexed by M, it is easy to see that $\{u_M(b), \ M > 1\}$ is a monotonically non-decreasing sequence of functions. Furthermore,

$$\lim_{M \to \infty} u_M(b) = K b^{C_H + \epsilon} 1_{\{b > b_0\}}, \ \forall \ b, b_0$$

where K is the positive constant in Equation (13). Invoking the MCT for the sequence $u_M(b)$, we have

$$\lim_{M \to \infty} \mathbb{E}_b\left[u_M(b)\right] = \mathbb{E}_b\left[\lim_{M \to \infty} u_M(b)\right] = \mathbb{E}_b\left[K b^{C_H + \epsilon}; \ b > b_0\right] = \infty.$$

Next, going back to (14) and taking M to infinity, we have

$$\lim_{M \to \infty} w_M = \lim_{M \to \infty}\left(\lim_{t \to \infty} \mathbb{E}\left[R_M(t)\right]\right) \geq \frac{1 - 3\kappa}{\mathbb{E}\left[T\right]} \lim_{M \to \infty} \mathbb{E}_b\left[u_M(b)\right] = \infty. \tag{15}$$

(ii) Recall that the steady-state queue backlog q_L is defined as the distributional limit of $q_L(t)$, as t becomes large. In other words, viewing $q_L(t)$ as a sequence of random variables indexed by t, we have $q_L(t) \Rightarrow q_L$, where "$\Rightarrow$" denotes convergence in distribution. Next, let us fix M, and view $R_M(t)$ as a sequence of random variables indexed by t. We have

$$R_M(t) \Rightarrow (M \wedge q_L)^{C_H - 1 + \epsilon}.$$

Theorem 25.12 in [1] asserts that when a sequence of random variables converges in distribution, the corresponding sequence of means also converges to the mean of the limiting random variable, as long as a technical condition called *uniform integrability* is satisfied. Since $R_M(t)$ is bounded above by $M^{C_H - 1 + \epsilon}$ for all t, uniform integrability is trivially satisfied, and we have

$$\lim_{t \to \infty} \mathbb{E}\left[R_M(t)\right] = \mathbb{E}\left[(M \wedge q_L)^{C_H - 1 + \epsilon}\right],$$

for each M. Thus,

$$w_M = \mathbb{E}\left[(M \wedge q_L)^{C_H - 1 + \epsilon}\right]. \tag{16}$$

(iii) Consider finally the term inside the expectation on the right hand side of Equation (16). When viewed as a sequence of random variables indexed by M, the term $(M \wedge q_L)^{C_H - 1 + \epsilon}$ represents a monotonically non-decreasing sequence of random variables. Furthermore,

$$\lim_{M \to \infty} (M \wedge q_L)^{C_H - 1 + \epsilon} = q_L^{C_H - 1 + \epsilon}.$$

Thus, another application of the MCT gives

$$\lim_{M \to \infty} \mathbb{E}\left[(M \wedge q_L)^{C_H - 1 + \epsilon} \right] = \mathbb{E}\left[q_L^{C_H - 1 + \epsilon} \right]. \tag{17}$$

Finally, combining (17), (16), and (15), we get

$$\mathbb{E}\left[q_L^{C_H - 1 + \epsilon} \right] = \lim_{M \to \infty} \mathbb{E}\left[(M \wedge q_L)^{C_H - 1 + \epsilon} \right] = \lim_{M \to \infty} w_M = \infty.$$

Proposition 5 is now proved. $\square$

APPENDIX B

PROOF OF PROPOSITION 7

For the renewal process considered above, consider the reward function $R_\gamma(t) = q_L(t)^{\gamma + \epsilon}$. Our aim is to show that the expected reward over the renewal interval is infinite, or

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{\gamma + \epsilon} \right] = \infty.$$

The key renewal theorem would then imply that $\lim_{t \to \infty} \mathbb{E}\left[q_L(t)^{\gamma + \epsilon} \right] = \infty$. We can finally appeal to a truncation argument to interchange the limit and the expectation, and obtain the desired result.

Defining $\mathcal{E}_H = \{ q_H(0) > 0, q_L(0) = 0 \}$, and proceeding as in the proof of Proposition 5,

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{\gamma + \epsilon} \right] \geq \mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{\gamma + \epsilon}; \ \mathcal{E}_H \right] = \mathbb{E}_b\left[\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{\gamma + \epsilon}; \ \mathcal{E}_H \Big| q_H(0) = b \right] \right].$$

In the last step above, we have iterated the expectation over the initial burst size b. The inner expectation above is a function of b; let us denote it by

$$g_\gamma(b) := \mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{\gamma + \epsilon}; \ \mathcal{E}_H \Big| q_H(0) = b \right].$$

Thus,

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{\gamma + \epsilon} \right] \geq \mathbb{E}_b\left[g_\gamma(b) \right] \geq \mathbb{E}_b\left[g_\gamma(b); \ b > b_0 \right], \ \forall b_0 \geq 1. \tag{18}$$

Since the above bound is true for any b_0, we can make b_0 as large as we want. We will make b_0 large enough for us to be able to invoke the law of large numbers several times in the rest of the proof.

At this point, we note that for the sake of a lower bound on the expected reward over the renewal interval, we can assume that the heavy queue receives no further arrivals after the initial burst. Under this assumption, we will next show that the light queue catches up with the heavy queue in $\Theta(b^{\alpha_H/\alpha_L})$ time slots. We first need to define what exactly we mean by 'catch-up'.

The catch-up time τ_c is defined as

$$\tau_c = \min\left\{ t > 0 \,\Big|\, q_L(t)^{\alpha_L/\alpha_H} \geq q_H(t) > 0 \right\}. \tag{19}$$

In words, the catch-up time is the first time after the arrival of the initial burst for which $q_L(\tau_c)^{\alpha_L/\alpha_H} \geq q_H(\tau_c)$. Note that the catch-up time need not always exist, even if $\mathcal{E}_H$ occurs[8]. However, we show that if the initial burst size is large, the catch-up time exists with high probability.

Indeed, let $b > b_0$ for large enough b_0, and suppose that a catch-up time does not exist. Let us consider the queue backlogs after the first $b - 1$ time slots, by which time the busy period could not have possibly ended. Since the light queue never catches up, the departure process from the light queue can be upper bounded by the number of exclusive slots. Thus, the light queue backlog at time $b - 1$ can be lower bounded as

$$q_L(b - 1) \geq \sum_{i=0}^{b-1} L(i) - S_L(i)(1 - S_H(i)).$$

Since catch-up has not occurred until time $b-1$, it follows that $q_L(b-1)^{\alpha_L/\alpha_H} < q_H(b-1) < b$. Thus, assuming that a catch-up time does not exist implies

$$\left(\sum_{i=0}^{b-1} L(i) - S_L(i)(1 - S_H(i)) \right)^{\alpha_L/\alpha_H} < b,$$

or equivalently,

$$\left(\frac{1}{b} \sum_{i=0}^{b-1} L(i) - S_L(i)(1 - S_H(i)) \right)^{\alpha_L/\alpha_H} < \frac{b}{b^{\alpha_L/\alpha_H}}.$$

When b is large, the weak LLN implies that the above event has a small probability. This is because the term inside the parentheses on the left is a sample average of random variables with positive mean. Thus, the non-occurrence of catch-up implies the occurrence of a small probability event. This implies that a catch-up time exists for large b with high probability[9].

---

[8]For example, the initial burst size might be small, and the system might empty again without the light queue ever receiving a single packet during the renewal interval.

[9]In this proof, when we state that an event occurs with high probability for large b, we mean the following: Given any $\kappa > 0$, there exists a large enough b_0 such that for all $b > b_0$, the event in question has probability greater than $1 - \kappa$. In a symmetric fashion, we can define a low probability event for large b as the complement of a high probability event.

Next, we show that τ_c is $\Theta(b^{\alpha_H/\alpha_L})$ with high probability. First, to obtain a lower bound on τ_c, define $\tau_1(b)$ as the unique positive solution to the equation

$$(\lambda_L \tau_1(b))^{\alpha_L} = (b - p_H \tau_1(b))^{\alpha_H}.$$

It is easy to see that $\tau_1(b) = \Omega(b^{\alpha_H/\alpha_L})$. Let us now bound the queue backlogs in the interval $0 \le t \le \lfloor \frac{\tau_1(b)}{2} \rfloor$. For the heavy queue,

$$q_H(t) \ge b - \sum_{i=0}^{\lfloor \frac{\tau_1(b)}{2} \rfloor} S_H(i) \ge b - (p_H + \delta) \lfloor \frac{\tau_1(b)}{2} \rfloor$$

with high probability for large b, where $\delta > 0$ can be chosen arbitrarily small. Similarly, for the light queue,

$$q_L(t) \le \sum_{i=0}^{\lfloor \frac{\tau_1(b)}{2} \rfloor} L(i) \le \lfloor \frac{\tau_1(b)}{2} \rfloor (\lambda_L + \delta)$$

with high probability for large b. Comparing the last two bounds, it is evident that

$$q_L(t)^{\alpha_L/\alpha_H} > q_H(t), \ 0 \le t \le \lfloor \frac{\tau_1(b)}{2} \rfloor,$$

for large b, with high probability. Thus, catch-up has not occurred by time $\lfloor \frac{\tau_1(b)}{2} \rfloor$, so that $\tau_c > \lfloor \frac{\tau_1(b)}{2} \rfloor$ with high probability for large b. Since $\tau_1(b) = \Omega(b^{\alpha_H/\alpha_L})$, it follows that τ_c is at least $\Omega(b^{\alpha_H/\alpha_L})$.

Second, to obtain an upper bound on the catch-up time, define

$$\tau_2(b) = \frac{(2b)^{\alpha_H/\alpha_L}}{\lambda_L - p_L(1 - p_H)}.$$

Suppose that catch-up has not occurred by time $\lceil \tau_2(b) \rceil$. Then, the departures from the light queue only occur during the exclusive slots of L. Thus,

$$q_L(\lceil \tau_2(b) \rceil) \ge \sum_{i=0}^{\lceil \tau_2(b) \rceil} L(i) - S_L(i)(1 - S_H(i)).$$

Since we assumed that catch-up has not occurred by time $\lceil \tau_2(b) \rceil$, we have $q_L(\lceil \tau_2(b) \rceil)^{\alpha_L/\alpha_H} < q_H(\lceil \tau_2(b) \rceil) \le b$. Therefore,

$$\left(\sum_{i=0}^{\lceil \tau_2(b) \rceil} L(i) - S_L(i)(1 - S_H(i)) \right)^{\alpha_L/\alpha_H} < b,$$

or equivalently,

$$\frac{1}{\lceil \tau_2(b) \rceil} \sum_{i=0}^{\lceil \tau_2(b) \rceil} L(i) - S_L(i)(1 - S_H(i)) < \frac{b^{\alpha_H/\alpha_L}}{\lceil \tau_2(b) \rceil} < \frac{\lambda_L - p_L(1 - p_H)}{2^{\alpha_H/\alpha_L}}.$$

By the weak LLN, the above event is of low probability when b is large. Therefore, we conclude that $\tau_c < \lceil \tau_2(b) \rceil$ with high probability when b is large.

We have so far shown that the light queue catches up with the heavy queue in a time scale of $\Theta(b^{\alpha_H/\alpha_L})$ with high probability. Therefore, it easily follows that $q_L(\tau_c) = \Theta(b^{\alpha_H/\alpha_L})$ and $q_H(\tau_c) = b - \Theta(b^{\alpha_H/\alpha_L})$ with high probability. We have now reached the core of the proof where we show that after τ_c, the light queue stays at $\Theta(b^{\alpha_H/\alpha_L})$ for $\Omega(b)$ time slots.

To this end, define σ_c as the first time after τ_c that the light queue backlog falls below $(q_H(\tau_c)/2)^{\alpha_H/\alpha_L}$. That is,

$$\sigma_c = \min\left\{ t > \tau_c \;\middle|\; q_L(t) < \left(\frac{q_H(\tau_c)}{2}\right)^{\alpha_H/\alpha_L} \right\}.$$

It is clear that σ_c is well defined when τ_c exists, since the system eventually empties.

With the intention of necessitating a low probability event, let us assume that

$$q_H(t) \geq \frac{3q_H(\tau_c)}{4}, \quad \text{for all } t \in [\tau_c, \sigma_c]. \tag{20}$$

Next, define

$$\omega_c = \max\left\{ \tau_c \leq t < \sigma_c \;\middle|\; q_L(t) \geq \left(\frac{3q_H(\tau_c)}{4}\right)^{\alpha_H/\alpha_L} \right\}.$$

In words, ω_c is the last time before σ_c that the light queue backlog exceeds $(3q_H(\tau_c)/4)^{\alpha_H/\alpha_L}$. Now, by the definition of ω_c and the assumption made in (20), it is clear that $q_L(t)^{\alpha_L/\alpha_H} < q_H(t)$ for $\omega_c < t \leq \sigma_c$. Thus, the departures that occur from the light queue during the interval $\omega_c < t \leq \sigma_c$ must necessarily occur during the exclusive slots of L. Therefore,

$$q_L(\sigma_c) = q_L(\omega_c) + \sum_{i=\omega_c+1}^{\sigma_c} L(i) - S_L(i)(1 - S_H(i)),$$

or equivalently,

$$\frac{1}{\sigma_c - \omega_c} \sum_{i=\omega_c+1}^{\sigma_c} L(i) - S_L(i)(1 - S_H(i)) = \frac{q_L(\sigma_c) - q_L(\omega_c)}{\sigma_c - \omega_c}.$$

This necessarily implies

$$\frac{1}{\sigma_c - \omega_c} \sum_{i=\omega_c+1}^{\sigma_c} L(i) - S_L(i)(1 - S_H(i)) < 0. \tag{21}$$

From the definition of σ_c and ω_c, it is clear that

$$\sigma_c - \omega_c > q_L(\omega_c) - q_L(\sigma_c) = \left(3^{\alpha_H/\alpha_L} - 2^{\alpha_H/\alpha_L}\right) \left(\frac{q_H(\tau_c)}{4}\right)^{\alpha_H/\alpha_L},$$

so that $\sigma_c - \omega_c$ is at least $\Omega(b^{\alpha_H/\alpha_L})$. Therefore, by the weak LLN, the event in (21) is a low probability event for large b.

What we have shown now is that the assumption in (20) implies the occurrence of a low probability event for large b. Therefore, the assumption (20) will be false with high probability when b is large. In other words, with high probability, there exists $t \in [\tau_c, \sigma_c]$ for which $q_H(t) < \frac{3q_H(\tau_c)}{4}$. In particular, this implies that $\sigma_c - \tau_c > \frac{q_H(\tau_c)}{4}$, with high probability for large b.

Next, since $q_H(\tau_c) = b - \Theta(b^{\alpha_H/\alpha_L})$ with high probability, we have $q_H(\tau_c) > b/2$ for large enough b. Thus, $\sigma_c - \tau_c > b/8$, with high probability, and for $\tau_c \leq t < \sigma_c$, the light queue backlog is lower bounded by

$$q_L(t) \geq \left(\frac{q_H(\tau_c)}{2} \right)^{\alpha_H/\alpha_L} > \left(\frac{b}{4} \right)^{\alpha_H/\alpha_L},$$

also with high probability. We have thus shown that after catch-up, the light queue backlog stays at $\Omega(b^{\alpha_H/\alpha_L})$ for $\Omega(b)$ slots, with high probability.

We can now return to (18) to finish the sequence of inequalities. In particular, let us choose b_0 large enough such that for $b > b_0$, the intersection of all the high probability events above has probability at least $1 - \kappa$, for some $\kappa > 0$. Then,

$$\mathbb{E}\left[\sum_{i=0}^{T} q_L(i)^{\gamma+\epsilon} \right] \geq \mathbb{E}_b\left[g_\gamma(b); \ b > b_0 \right] \geq (1-\kappa)\mathbb{E}_b\left[\frac{b}{8} \cdot \left(\frac{b}{4} \right)^{\alpha_H/\alpha_L(\gamma+\epsilon)}; \ b > b_0 \right]$$

$$= K_1 \mathbb{E}_b\left[b \cdot b^{C_H - 1 + \epsilon \alpha_H/\alpha_L}; \ b > b_0 \right] = \infty,$$

since the burst size b has tail coefficient C_H. The key renewal theorem would then imply that

$$\lim_{t \to \infty} \mathbb{E}\left[q_L(t)^{\gamma+\epsilon} \right] = \infty, \ \text{for all} \ \epsilon > 0.$$

We can finally invoke a truncation argument similar to the one in Proposition 5 to interchange the limit and the expectation. Thus, for the steady-state backlog q_L, we have $\mathbb{E}\left[q_L^{\gamma+\epsilon} \right] = \infty, \ \text{for all} \ \epsilon > 0.$ $\square$

REFERENCES

[1] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, New York, 1995.

[2] A. Brzezinski and E. Modiano, "Dynamic reconfiguration and routing algorithms for IP-over-WDM networks with stochastic traffic," *Journal of Lightwave Technology*, vol. 23, no. 10, p. 3188, 2005.

[3] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 411–424, 2005.

[4] R. Gallager, *Discrete stochastic processes*. Kluwer, 1996.

[5] A. Ganesh, N. O'Connell, and D. Wischik, *Big queues*. Springer Verlag, 2004.

[6] A. Ganti, E. Modiano, and J. Tsitsiklis, "Optimal transmission scheduling in symmetric communication models with intermittent connectivity," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 998–1008, 2007.

[7] K. Jagannathan, "Asymptotic performance of queue-length-based network control policies," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2010.

[8] K. Jagannathan, M. Markakis, E. Modiano, and J. Tsitsiklis, "Throughput Optimal Scheduling in the presence of Heavy-Tailed Traffic," in *Forty-eigth Annual Allerton Conference on Communication, Control, and Computing, Monticello IL*, 2010.

[9] K. Jagannathan, M. Markakis, E. Modiano, and J. N. Tsitsiklis, "Queue-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.

[10] L. Le, K. Jagannathan, and E. Modiano, "Delay analysis of maximum weight scheduling in wireless ad hoc networks," *Proc. of IEEE CISS*, pp. 389–394, 2009.

[11] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on networking*, vol. 2, no. 1, pp. 1–15, 1994.

[12] M. Markakis, E. Modiano, and J. Tsitsiklis, "Scheduling policies for single-hop networks with heavy-tailed traffic," in *Allerton Conference, Monticello IL*, 2009.

[13] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1260–1267, 1999.

[14] J. Nair, K. Jagannathan, and A. Wierman, "When heavy-tailed and light-tailed flows compete: the response time tail under generalized max-weight scheduling," in *Proc. IEEE INFOCOM*, 2013.

[15] M. Neely, "Delay analysis for max weight opportunistic scheduling in wireless systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 9, pp. 2137–2150, 2009.

[16] M. Neely, E. Modiano, and C. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," in *IEEE INFOCOM*, vol. 3, 2002.

[17] ——, "Dynamic power allocation and routing for time varying wireless networks," in *IEEE INFOCOM*, vol. 1, 2003.

[18] A. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: Large deviations and optimality," *Annals of Applied Probability*, vol. 11, no. 1, pp. 1–48, 2001.

[19] V. Subramanian, "Large deviations of max-weight scheduling policies on convex rate regions," *Proceedings of ITA*, 2008.

[20] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[21] ——, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 466–478, 1993.