When Heavy-Tailed and Light-Tailed Flows Compete: The Response Time Tail Under Generalized Max-Weight Scheduling

Jayakrishnan Nair, Krishna Jagannathan, and Adam Wierman

***Abstract*—This paper focuses on the design and analysis of scheduling policies for multi-class queues, such as those found in wireless networks and high-speed switches. In this context, we study the response-time tail under generalized max-weight policies in settings where the traffic flows are highly asymmetric. Specifically, we consider a setting where a bursty flow, modeled using heavy-tailed statistics, competes with a more benign, light-tailed flow. In this setting, we prove that classical max-weight scheduling, which is known to be throughput optimal, results in the light-tailed flow having heavy-tailed response times. However, we show that via a careful design of inter-queue scheduling policy (from the class of generalized max-weight policies) and intra-queue scheduling policies, it is possible to maintain throughput optimality, and guarantee light-tailed delays for the light-tailed flow, without affecting the response-time tail for the heavy-tailed flow.**

***Index Terms*—First come first served, heavy-tailed traffic, large deviations, last come first served, light-tailed traffic, maximum weight scheduling, response time tail, stability.**

I. Introduction

The task of scheduling conflicting links is central to a variety of networking settings, such as wireless networks, optical networks and high-speed switches. As a result, there is a large literature studying scheduling policies in these contexts, most of which is based on the maximum-weight (max-weight) scheduling framework proposed by Tassiulas and Ephremides in [2], [3]. At this point, there is a substantial body of literature devoted to the analysis and application of the max-weight policy and its variants; for example, see [4]–[10].

Traditionally, the focus of research on max-weight scheduling has been on understanding its 'stability region', i.e., the set of input rates that can be supported. Notably, max-weight has been shown to be 'throughput optimal' in very general settings, i.e., it has the largest possible stability region among all scheduling policies [2], [3], [10]. In other words, if there exists any scheduling policy that can keep the queueing network stable under a given model of traffic arrival statistics, the max-weight policy can stabilize the system.

Although throughput is an important first-order performance metric, a more discerning metric is the *response time*, a.k.a., sojourn time or delay. Indeed, from the standpoint of the applications sending/receiving information, ensuring small, predictable response times is crucial. Although the stability region and throughput optimality properties of the max-weight framework are well studied, the literature on the delay performance is relatively small. Average delay bounds are derived using Lyapunov drift techniques in some works (for example, see [10]); however, these are quite loose in general. Tighter delay bounds have been established recently; see, for example, [11], [12].

In general, results about the response time of max-weight policies, such as those above, tend to indicate that max-weight policies perform well in symmetric traffic settings [2], [13], [14]. This is primarily due to the tendency of these policies to 'balance out' the queues in the system, by preferentially serving longer queues. For example, [2] contains a strong sample path optimality result for queue backlogs under stochastically symmetric traffic to parallel queues; this is generalized in [13].

On the other hand, the traffic flows encountered in practice tend to be highly asymmetric, with a wide range of variability or burstiness. Indeed, in the context of communication networks, certain bursty traffic flows may be well modeled using heavy-tailed arrival processes, and the more benign ones better modeled using light-tailed processes. For example, an internet user might generate occasional file download requests with highly variable file sizes, that can be modeled as being heavy-tailed. However, routine webpage loading and email traffic are likely to be far less variable, and thus are better modeled as being light-tailed. In order to capture the interaction between heterogeneous traffic sources in a queueing network, multi-class queueing models with a mix of heavy-tailed and light-tailed traffic sources have been studied [15]–[18]. An important paper in this category is [15], where the interaction between light and heavy-tailed traffic flows under generalized processor sharing (GPS) is studied. Another example is [16],

Manuscript received March 17, 2013; revised February 19, 2014; accepted January 19, 2015; approved by IEEE/ACM Transactions on Networking Editor I. Keslassy. Date of publication March 30, 2015; date of current version April 14, 2016. The work of J. Nair and A. Wierman was supported in part by the NSF through grant CNS 0846025 and NetSE grant CNS 0911041, in part by the ARO through MURI grant W911NF-08-1-0233, and in part by Bell Labs, Alcatel-Lucent. The work of J. Nair was also supported in part by an NWO VIDI grant. The work of K. Jagannathan was supported in part by ARO MURI grant W911NF-08-1-0238 and in part by the Indo UK Advanced Technology Center (IUATC).

J. Nair is with the Department of Electrical Engineering, IIT Bombay, Maharashtra 400076, India (e-mail: ujk@caltech.edu).

K. Jagannathan is with the Department of Electrical Engineering, IIT Madras, Tamil Nadu 600036, India.

A. Wierman is with the Computing and Mathematical Sciences Department, California Institute of Technology, Pasadena, CA 91125 USA.

Digital Object Identifier 10.1109/TNET.2015.2415874

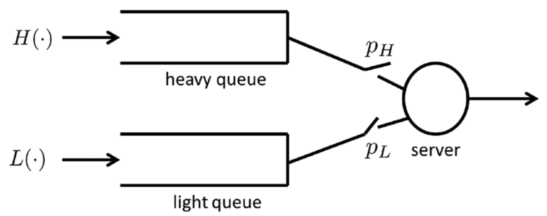


Fig. 1. A network consisting of two parallel queues, with one of them fed with heavy-tailed traffic. The channels connecting the queues to the server are unreliable ON/OFF links.

where the authors obtain the asymptotic workload behavior under a general coupled-queues framework, which includes GPS as a special case.

In summary, on the one hand, max-weight policies are throughput optimal and often provide good response times when the traffic is largely symmetric. On the other hand, the interaction between bursty and benign traffic sources is well studied within multi-class queueing and GPS frameworks, but these policies are not throughput optimal.

Contributions of this Paper: The goal of this paper is to fill this gap by studying response times under max-weight policies when traffic is highly asymmetric. The first steps towards filling this gap have been provided by the recent work of Markakis *et al.* in [19] and Jagannathan *et al.* [20], [21], which analyze a scenario where heavy-tailed and light-tailed flows interact through a generalized max-weight policy. Our present paper builds on these papers; in particular, our model is the same as in [20]. However, the focus of the above papers is on *queue length* asymptotics under different throughput optimal policies, while in this paper, we analyze the distribution of *response times* experienced by the heavy and light-tailed flows.

More specifically, in this paper, we consider a stylized setting where the traffic asymmetry is extreme. We consider a system consisting of two traffic classes contending for service from a single server, where one class is heavy-tailed, and the other is light-tailed (see Fig. 1). Both classes experience a time varying connectivity with the server, and the server can serve a single packet from a connected queue in each slot. Note that this model captures a wireless uplink/downlink scenario with two nodes communicating with an access point or base station via fading channels. For this queueing system, we study the tail of the (stationary) response-time distribution that each traffic class experiences under generalized max-weight policies.

In this context, there are two scheduling decisions: the *inter-queue* scheduling and the *intra-queue* scheduling. The inter-queue scheduling policy determines which queue to serve in each slot, whereas the intra-queue scheduling policies specify which waiting packet to serve from the queue selected for service by the inter-queue scheduling policy.

The *first contribution* of this paper is to prove that the classical max-weight policy, which serves the longest connected queue in each slot, causes the light-tailed flow to experience heavy-tailed response times. This means that the classical max-weight policy, while being throughput optimal, severely throttles (starves) the light-tailed flow. Thus, while max-weight performs well in symmetric settings, it can have poor performance in asymmetric settings. Intuitively, this is because the max-weight policy starves the light-tailed flow of service for a long period of time when the heavy-tailed flow generates its (frequent) large bursts.

The *second contribution* of this paper is to show that it is possible to design a throughput optimal scheduling policy that avoids the problems experienced by the classical max-weight policy. In particular, we present a policy that provably guarantees light-tailed response times for the light-tailed flow. Importantly, our results suggest that the response-time tail for the heavy-tailed flow remains unaffected; we prove this formally for the special case in which both queues are always connected to the server.

Our policy design entails a careful choice of the inter-queue scheduling policy, as well as intra-queue scheduling policies. Our inter-queue policy is the so called 'log-max-weight policy', which belongs to the class generalized max-weight policies [7] and awards a significant priority to the light-tailed flow, while maintaining throughput optimality. Our intra-queue policy differs between the heavy-tailed and the light-tailed queues: within the heavy-tailed queue, Preemptive-Last-Come-First-Served (PLCFS) is used, while within the light-tailed queue, First-Come-First-Served (FCFS) is used.

Our analysis provides a clear insight into the intricate interplay between the intra-queue and inter-queue scheduling policies. Indeed, our results reveal that even with a good inter-queue scheduling policy, the correct choice of intra-queue scheduling policies is crucial in order to obtain good response-time tail behavior. In fact, the difference in response times between two intra-queue policies can be significantly larger under generalized max-weight inter-queue scheduling than in a single server queue.

Finally, it is worth commenting that in attaining the results described above, we also settle an an open question in [21, pp. 171] regarding the asymptotics of log-max-weight scheduling. In particular, we prove that under log-max-weight scheduling, the (stationary) queue length distribution corresponding to the light-tailed queue is light-tailed (Theorem 8), via a novel application of Lyapunov bounds from [7].

II. Model and Preliminaries

A. System Model

Our goal is to study multi-class queues in a setting where the traffic flows are highly asymmetric. To that end, we consider a simple model where the asymmetry is extreme. In particular, we consider a scenario where two parallel queues contend for service from a single server. One of the queues sees a heavy-tailed arrival process, whereas the other sees a light-tailed arrival process. We refer to the former queue as the heavy queue, and the latter queue as the light queue.

Each queue experiences a stochastically time varying connectivity with the server. Fig. 1 provides an illustration of our setup. Time is slotted, and in each slot, the server can provide a single unit of service to a connected queue. Henceforth, we refer to this unit of service as a packet, and say the server can process a single packet from a connected queue in each slot. Let t denote the time index.

In each slot, a job, comprising a burst of packets, can arrive stochastically into each queue. Let $H(t)$ and $L(t)$ denote, respectively, the size of the job (in number of packets) arriving into the heavy queue and the light queue in time slot t. We adopt the convention that the size of the incoming job is zero if there is no arrival in a slot.

Our stochastic model for the arrival processes is the following. The sequences $L(\cdot)$ and $H(\cdot)$ are i.i.d. across time slots, and independent of one another. The random variable $L(t)$ is light-tailed, and the random variable $H(t)$ is heavy-tailed. Specifically, we assume that $H(t)$ is regularly varying with index $\theta_H > 1$.[1] Let $\lambda_H := \mathbb{E}[H(t)]$ and $\lambda_L := \mathbb{E}[L(t)]$ denote the mean arrival rates into the heavy queue and the light queue, respectively.

Next, we describe the stochastic model for the connectivity of each queue with the server. The connectivity of the heavy queue and the light queue are described, respectively, by Bernoulli sequences $\{\eta_H(t)\}$ and $\{\eta_L(t)\}$. $\eta_H(t), \eta_L(t) \in \{0,1\}$, with a value of 1 indicating that the corresponding queue is connected to the server in time slot t. We assume that the sequences $\{\eta_H(t)\}$ and $\{\eta_L(t)\}$ are mutually independent and independent of the arrival processes. Let $p_H := P(\eta_H(t) = 1)$ and $p_L := P(\eta_L(t) = 1)$ denote, respectively, the probabilities that the heavy queue and the light queue are connected to the server in each time slot. We assume that $p_L, p_H > 0$. We refer to the special case of our model in which the two queues are always connected to the server, i.e., $p_L = p_H = 1$, as the *wireline scenario*. For technical reasons, we exclude from consideration the scenario where only one of the queues is always connected to the server, i.e., we exclude the cases $p_L = 1,\ p_H \in (0,1)$ and $p_H = 1,\ p_L \in (0,1)$. Finally, we assume that the server can detect the connectivity state of both queues, as well as the queue size (in number of packets) of a connected queue in each slot. Note that our model captures an uplink/downlink setting with two wireless nodes connected to a base station or access point via independent fading channels.

Let $q_H(t)$ and $q_L(t)$ denote, respectively, the lengths (in number of packets) of the heavy queue and the light queue in the beginning of time slot t. The queue lengths evolve as follows:

$$q_H(t+1) = H(t) + q_H(t) - \mathbf{1}_{\{\text{heavy queue got service in slot t}\}}$$
$$q_L(t+1) = L(t) + q_L(t) - \mathbf{1}_{\{\text{light queue got service in slot t}\}}.$$

If both queues are connected to the server in a certain slot, the scheduling policy determines which queue will receive service. If only one of the queues is connected to the server in a certain slot, then that queue receives service if it has any waiting packets. We refer to such slots as exclusive slots. We use q_H and q_L to denote, respectively, the stationary queue lengths of the heavy queue and the light queue. We use V_H to denote the steady state response time experienced by a job in the heavy queue, and V_L to denote the steady state response time experienced by a job in the light queue.

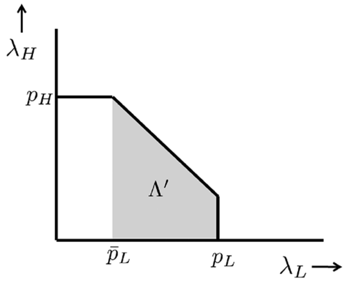


Fig. 2. Stability region Λ is the pentagonal region above. The subset Λ' of interest is shaded.

B. Stability Region

The *stability region* for the queueing system defined above, i.e., the set Λ of (λ_H, λ_L) pairs that are stabilizable, is well understood. It follows from [2] that

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

The stability region is visualized in Fig. 2. We seek scheduling policies that are *throughput optimal*, i.e., policies that stabilize the queueing system over the entire stability region.

Let $\bar{p}_L := p_L(1 - p_H)$. Note that $\bar{p}_L$ is the probability that only the light queue is connected to the server in a slot, i.e., the probability that a slot is exclusive to the light queue. If $\lambda_L < \bar{p}_L$, then the arrivals into the light queue can be stably supported by just exclusive slots, implying the light queue essentially does not need to compete for service with the heavy queue. This case is uninteresting when analyzing the light queue, since the response-time distribution is guaranteed to be light-tailed, irrespective of the inter-queue or intra-queue scheduling policy. For the same reason, the case $\lambda_H = 0$ is uninteresting. Therefore, when studying the response-time distribution in the light queue, we restrict our attention to the subset Λ' of the stability region over which $\lambda_L > \bar{p}_L$, and $\lambda_H > 0$. The set Λ' is depicted as the shaded region in Fig. 2. Note that in the wireline scenario, $\bar{p}_L = 0$, and Λ' is simply the interior of the stability region.

C. Scheduling

We decouple the scheduling design as follows. The *inter-queue scheduling policy* determines which queue to serve in each slot, given the connectivity state and length (in number of packets) of each queue. In the queue selected for service by the inter-queue policy, the *intra-queue scheduling policy* determines which packet to serve in that slot, given the full state of the queue. We consider a variety of possible policies, described below, for each.

Recall that we have two performance goals for scheduler design: (i) throughput optimality, and (ii) good response-time tail behavior. Note that the stability of the queueing system depends solely on the inter-queue scheduling policy, since the evolution of the queue lengths is insensitive to the intra-queue scheduling policy. However, the response-time distribution is highly dependent on the intra-queue scheduling policy.

1) Inter-Queue Scheduling: Given that the inter-queue scheduling policy completely determines the stability of the

[1] We formally define light-tailed and regularly varying distributions in Section II-D.

system, it is crucial to use policies that are throughput optimal. This motivates us to consider generalized max-weight policies [7]. In particular, our focus is on two such policies:

Max-Weight-α Scheduling: The max-weight-α policy [19], [21] is a generalization of the classical max-weight policy, and is characterized by two positive parameters α_L and α_H. In each slot, the max-weight-α policy serves the queue that wins the comparison

$$q_L(t)^{\alpha_L}\eta_L(t) \gtreqless q_H(t)^{\alpha_H}\eta_H(t). \tag{1}$$

Ties may be broken arbitrarily, but we assume for concreteness that ties are broken in favor of the light queue. Note that when $\alpha_L = \alpha_H$, the max-weight-α policy is identical to the classical max-weight policy. The throughput optimality of this policy follows easily from Theorem 1 in [7].

The parameters α_L and α_H determine the relative priorities of the two queues. Since we will be interested in the scenario where the light queue receives a higher priority than the heavy queue, we focus on the case $\alpha_L \geq \alpha_H$. Moreover, it is easy to see that we may set $\alpha_H = 1$ without loss of generality. Indeed, the comparison in (1) is equivalent to the comparison

$$q_L(t)^{\alpha_L/\alpha_H}\eta_L(t) \gtreqless q_H(t)\eta_H(t).$$

Accordingly, we focus on the range of parameters satisfying $\alpha_L \geq \alpha_H = 1$. Note that a higher value of α_L implies a higher priority for the light queue.

Log-Max-Weight Scheduling: The log-max-weight policy [21] is defined as follows. In each slot t, it serves the queue that wins the comparison

$$q_L(t)\eta_L(t) \gtreqless \log\left(1 + q_H(t)\right)\eta_H(t). \tag{2}$$

As before, we assume for concreteness that ties are broken in favor of the light queue. The throughput optimality of this policy once again follows easily from Theorem 1 in [7].

The log-max-weight policy awards an even higher degree of priority to the light queue than the max-weight-α policy. Note that in order to determine which queue to serve in a slot, the max-weight-α policy compares $q_H(t)$ with $q_L(t)^{\alpha_L}$, whereas the log-max-weight policy compares $q_H(t)$ with $e^{q_L(t)} - 1$.

2) Intra-Queue Scheduling: While intra-queue scheduling does not impact the stability of the system (as long as the policies considered are work-conserving), the intra-queue scheduling policy does have a significant impact on the response-time distribution. In this paper, we focus on two candidate policies for intra-queue scheduling: First-Come-First-Served (FCFS) and Preemptive-Last-Come-First-Served (PLCFS).

While other policies could also be considered, the choice of these policies is motivated by a few important factors. First, FCFS is the most commonly assumed intra-queue policy in the literature on max-weight scheduling. Second, there have been suggestions recently that using PLCFS as the intra-queue scheduling policy can improve the delay-performance of max-weight policies [22]. Third, in a single server queue, it is known that the response-time tail under FCFS is optimal when job sizes are light-tailed, while the response-time tail under PLCFS is optimal (up to a constant) when job sizes are heavy-tailed (see [23]).

D. Heavy-Tailed Distributions: Definitions and Properties

In this section, we give relevant definitions and preliminaries related to heavy-tailed distributions.

For any non-negative random variable X, we use F_X to denote its distribution function (d.f.), i.e., $F_X(x) := P(X \leq x)$, and $\bar{F}_X$ to denote its tail distribution function, i.e., $\bar{F}_X(x) := P(X > x)$. The random variable X (or its d.f. F_X) is said to be *heavy-tailed* if

$$\limsup_{x\to\infty} \frac{\bar{F}_X(x)}{e^{-\phi x}} = \infty \qquad \forall\, \phi > 0.$$

Conversely, X (or its d.f. F_X) is said to be *light-tailed* if it is not heavy-tailed, i.e., if there exists $\phi > 0$ such that

$$\lim_{x\to\infty} \frac{\bar{F}_X(x)}{e^{-\phi x}} = 0.$$

Intuitively, a d.f. is heavy-tailed if its tail is asymptotically heavier than that of any exponential distribution.

An important characterization of heavy-tailed distributions that we make use of in our analysis is the following (see Theorem 2.6 in [24]). For any non-negative random variable X, define $\Psi_X(x) := -\log \bar{F}_X(x)/x$.

Lemma 1: Suppose X is non-negative random variable. Then X is heavy-tailed if and only if

$$\liminf_{x\to\infty} \Psi_X(x) = 0.$$

From a modeling standpoint, an important subclass of heavy-tailed distributions is the class of regularly varying distributions, which is a generalization of the class of Pareto distributions [25]. Formally, a random variable X (or its d.f. F_X) is said to be *regularly varying* with index $\theta > 0$ (denoted $X \in \mathcal{RV}(\theta)$) if $P(X > x) = x^{-\theta}L(x)$, where $L(x)$ is a slowly varying function, i.e., $L(x)$ satisfies $\lim_{x\to\infty}(L(xy)/L(x)) = 1\ \forall\ y > 0$. Recall that our model assumes that $H(t) \in \mathcal{RV}(\theta_H)$.

Our focus in this paper is on understanding the (logarithmic) asymptotic behavior of the response-time tail. To study this for a heavy-tailed X, we use its *tail index*, defined as

$$\Gamma(X) := \lim_{x\to\infty} -\frac{\log P(X > x)}{\log(x)},$$

when the limit exists. The tail index is useful for describing the asymptotic tail behavior of distributions that exhibit a roughly 'power-law' tail, such as regularly varying distributions. In particular, if $X \in \mathcal{RV}(\theta)$, then $\Gamma(X) = \theta$ [26, Prop. 2.6]. It is easy to check that if $\Gamma(X) < \infty$, then X is heavy-tailed. Moreover, it can be shown that

(i) if $\Gamma(X) > 0$, then $\mathbb{E}[X^\beta] < \infty$ for $0 \leq \beta < \Gamma(X)$,

(ii) if $\Gamma(X) < \infty$, then $\mathbb{E}[X^\beta] = \infty$ for $\beta > \Gamma(X)$.

Finally, note that a smaller value of tail index implies a 'heavier' tail.

To give a lower bound on the tail of a heavy-tailed random variable X, we use

$$\bar{\Gamma}(X) := \limsup_{x\to\infty} -\frac{\log P(X > x)}{\log(x)}.$$

It is easy to check that if $\bar{\Gamma}(X) < \infty$, then X is heavy-tailed. Moreover, if $\bar{\Gamma}(X) < \infty$, then $\mathbb{E}[X^\beta] = \infty$ for $\beta > \bar{\Gamma}(X)$.

TABLE I
SUMMARY OF MAIN RESULTS. $V_{i,\pi}$ IS THE STATIONARY RESPONSE TIME IN QUEUE i UNDER INTRA-QUEUE SCHEDULING POLICY π

	Light queue	Heavy queue
Max-weight-α $(\alpha_L \geq \alpha_H = 1)$	$\Gamma(V_{L,FCFS}) = \alpha_L(\theta_H - 1)$ (Thm. 2) $\bar{\Gamma}(V_{L,PLCFS}) \leq \theta_H - 1/\alpha_L$ (Thm. 3)	$\Gamma(V_{H,FCFS}) = \theta_H - 1$ (Thm. 4) $\Gamma(V_{H,PLCFS}) = \theta_H$ (Thm. 5) (wireline only; for $\alpha_L > \frac{\theta_H}{\theta_H - 1}$)
Log-max-weight	$V_{L,FCFS}$ is light-tailed (Thm. 6) $V_{L,PLCFS}$ is heavy-tailed (Thm. 7)	$\Gamma(V_{H,FCFS}) = \theta_H - 1$ (Thm. 9) (wireline only) $\Gamma(V_{H,PLCFS}) = \theta_H$ (Thm. 10) (wireline only)

III. RESULTS

The multi-class queueing model in the previous section involves two highly asymmetric traffic classes, one heavy-tailed, and one light-tailed. Our goal now is to understand how max-weight scheduling and its variants perform under such an extreme form of asymmetry. We begin by considering the most well studied subclass of generalized max-weight policies: max-weight-α policies [19], which includes the classical max-weight policy as a special case. We then consider the log-max-weight policy. Our main results are summarized in Table I.

Recall that both of these classes of inter-queue policies ensure throughput stability regardless of the intra-queue policy used. Therefore, our results focus on the response-time tail. Importantly, for this metric, our results highlight that the choice of intra-queue scheduling is crucial.

A. Max-Weight-α Scheduling

In this section, we present our results on the tail behavior of the (stationary) response-time distribution in the heavy queue and light queue under the max-weight-α inter-queue scheduling policy. We being by focusing on the light-queue.

The Performance of the Light Queue: Our first result is the following upper bound on the response-time tail index for the light queue under max-weight-α inter-queue scheduling, and any intra-queue scheduling policy in the light queue.

Theorem 1: Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under the max-weight-α scheduling policy between queues with $\alpha_L \geq \alpha_H = 1$

$$\bar{\Gamma}(V_L) = \limsup_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} \leq \alpha_L \theta_H - 1$$

for any intra-queue scheduling policy in the light queue.

Theorem 1 states that under max-weight-α scheduling between queues, $\bar{\Gamma}(V_L) < \infty$, which implies that *the light queue sees heavy-tailed response times*, irrespective of the intra-queue scheduling policy. This means that although max-weight-α scheduling is throughput optimal, it severely throttles the light queue. Note that this includes the classical max-weight policy as a special case. Intuitively, this poor performance is the result of (frequent) large arrivals into the heavy queue starving the light queue of service for a long time.

However, it is important to note that the upper bound on the response-time tail index given by Theorem 1 is an increasing function of α_L, approaching ∞ as $\alpha_L \to \infty$. This suggests the possibility of achieving an arbitrarily large response-time tail index for the light queue (recall that a larger tail index implies a lighter tail) by setting α_L large enough, i.e., by awarding the light queue sufficiently high priority. Theorems 2 and 3 below imply that this is indeed the case, so long as the intra-queue policy in the light queue is chosen appropriately. Intuitively, a larger value of α_L makes the interval of service starvation of the light queue following the arrival of a large job into the heavy queue shorter, thus improving the response-time tail.

Theorem 2: Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under max-weight-α scheduling between queues with $\alpha_L > \alpha_H = 1$, and First-Come-First-Served scheduling within the light queue

$$\Gamma(V_L) = \lim_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} = \alpha_L(\theta_H - 1).$$

Theorem 3: Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under max-weight-α scheduling between queues with $\alpha_L > \alpha_H = 1$, and Preemptive-Last-Come-First-Served scheduling within the light queue

$$\bar{\Gamma}(V_L) = \limsup_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} \leq \theta_H - \frac{1}{\alpha_L}.$$

Theorem 2 states that with FCFS scheduling within the light queue, the response-time tail index increases linearly with α_L. This means that while the response-time distribution in the light queue remains heavy-tailed for all α_L, its tail index can be made arbitrarily large by setting α_L to a large enough value, i.e., by giving the light queue sufficient priority. In contrast, Theorem 3 states that under PLCFS scheduling in the light queue, the tail index remains bounded above by θ_H for all values of α_L. This highlights the importance of choosing the correct intra-queue scheduling policy in order to exploit the priority awarded to it by the inter-queue scheduling policy.

The Performance of the Heavy Queue: Next, we turn to the response-time tail in the heavy queue under max-weight-α inter-queue scheduling. The following theorems summarize our results for FCFS and PLCFS intra-queue scheduling in the heavy queue.

Theorem 4: Under max-weight-α scheduling between queues with $\alpha_L \geq \alpha_H = 1$, and First-Come-First-Served scheduling within the heavy queue

$$\Gamma(V_H) = \lim_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} = \theta_H - 1.$$

Theorem 5: In the wireline scenario, under max-weight-α scheduling policy between queues with $\alpha_L > \theta_H/(\theta_H - 1)$ and $\alpha_H = 1$, and Preemptive-Last-Come-First-Served scheduling within the heavy queue,

$$\Gamma(V_H) = \lim_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} = \theta_H. \quad (3)$$

Theorem 4 implies that with FCFS scheduling within the heavy queue, the response-time tail index is insensitive to α_L; i.e., it is insensitive to the level of relative priority awarded to the light queue. Moreover, the response-time tail index is the same as it would be in an isolated $Geo/GI/1$ queue with the same arrival process as the heavy queue.[2]

With PLCFS scheduling within the heavy queue, we are only able to analyze the wireline scenario, when the inter-queue priority to the light queue being sufficiently high (specifically, $\alpha_L > \theta_H/(\theta_H - 1)$). For this case, Theorem 5 implies as before that the response-time tail index for the heavy queue is insensitive to α_L, and is the same as it would be in an isolated $Geo/GI/1$ queue with the same arrival process.[2] Furthermore, this response-time tail index is optimal, since the response-time tail index is bounded above by the tail index of the job size distribution (i.e., θ_H). We conjecture that Equation (3) holds even in our general ‘wireless’ scenario, in which the two queues have a stochastic connectivity with the server.[3]

To summarize, under max-weight-α scheduling, the light queue necessarily experiences heavy-tailed response times. However, by setting α_L large enough, i.e., by awarding sufficiently high priority to the light queue, its response-time tail index can be made arbitrarily large, with the correct choice of intra-queue scheduling policy. Further, our results suggest that the response-time tail index of the heavy queue is unaffected in this process, and behaves like the response-time tail index in an isolated $Geo/GI/1$ queue (with the same arrival process).

Ultimately however, from a fairness standpoint, it is desirable that response times in the light queue are light-tailed. Since the level of priority awarded to the light queue by the max-weight-α policy is insufficient for this to happen, we now analyze the log-max-weight inter-queue policy, which awards an even higher degree of relative priority to the light queue.

B. Log-Max-Weight Scheduling

In this section, we study the tail behavior of the (stationary) response-time distribution in the light queue and the heavy queue under the log-max-weight inter-queue scheduling policy.

The Performance of the Light Queue: Our main result in this section is that under log-max-weight scheduling between queues, and FCFS scheduling within the light queue, the light queue experiences light-tailed response times.

Theorem 6: Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under log-max-weight scheduling between queues, and First-Come-First-Served scheduling within the light queue, V_L is light-tailed.

The above theorem implies that *the log-max-weight policy indeed provides sufficient priority to the light queue to make its response-time distribution light-tailed*. However, for this to happen, the intra-queue scheduling policy cannot be chosen arbitrarily. In fact, as the following theorem shows, with PLCFS scheduling within the light queue, its response-time distribution remains heavy-tailed.

Theorem 7: Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under log-max-weight scheduling between queues, and Preemptive-Last-Come-First-Served scheduling within the light queue, V_L is heavy-tailed.

This extreme contrast between the two policies highlights once again the importance of correctly choosing the intra-queue scheduling policy to exploit the priority awarded to the light queue by the inter-queue scheduling policy. Theorems 6 and 7 demonstrate a remarkable phenomenon: with the same service process for the light queue, one intra-queue scheduling discipline results in heavy-tailed response times, whereas another leads to light-tailed response times. In the context of the $Geo/G/1$ queue, the impact of the intra-queue policy is nowhere near this extreme, which highlights how crucial the choice is for the multi-queue setting.

The proof of Theorem 6 relies crucially on the following.

Theorem 8: Under log-max-weight scheduling between queues, q_L is light-tailed.

This statement was originally conjectured in [21], but proved only for the wireline scenario. In Section IV, we give a novel proof of Theorem 8 based on Lyapunov arguments.

The Performance of the Heavy Queue: Under log-max-weight inter-queue scheduling, we are only able to analyze the response-time tail for the heavy queue in the wireline scenario. For this case, we prove that with both FCFS and PLCFS intra-queue scheduling, the response-time distribution has the same tail index as in an isolated $Geo/GI/1$ queue with the same arrival process. These results show that in the wireline scenario, the response-time tail index is unaffected by the priority given to the light queue by the log-max-weight policy. We conjecture that the same is true in our general ‘wireless’ model.[3] The following theorems summarize our results.

Theorem 9: In the wireline scenario, under log-max-weight scheduling between queues, and First-Come-First-Served scheduling within the heavy queue

$$\lim_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} = \theta_H - 1.$$

Theorem 10: In the wireline scenario, under log-max-weight scheduling between queues, and Preemptive-Last-Come-First-Served scheduling within the heavy queue,

$$\lim_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} = \theta_H.$$

To summarize, our results show that it is possible to achieve light-tailed response times in the light queue using log-max-weight inter-queue scheduling. In other words, it is possible

[2] In a $Geo/GI/1$ queue with the same arrival process as the heavy queue, it is well known that the response-time tail index equals $\theta_H - 1$ under FCFS scheduling, and θ_H under PLCFS scheduling (for example, see [23]).

[3] The extension to the ‘wireless’ case is made difficult by the fact that the busy period tail behavior is unknown for this case. On the other hand, in the wireline scenario, the busy period behaves identically to busy periods in a $Geo/GI/1$ queue which sees the combined arrival processes of the heavy and the light queue in our model; this busy period is well understood.

to design inter-queue and intra-queue scheduling policies for our system such that we maintain throughput optimality, and achieve light-tailed delays for the light queue. Importantly, our results suggest that this can be done without affecting the response-time tail index for the heavy queue.

IV. Proofs

This section is devoted to proofs of the results presented in Section III.

First, we introduce some notation that is used heavily in our proofs. For functions $\varphi(x)$ and $\xi(x)$, the notation $\varphi(x) \sim \xi(x)$ means $\lim_{x\to\infty}(\varphi(x)/\xi(x)) = 1$. For $t_1, t_2 \in \mathbb{N}$, $A_L(t_1, t_2) := \sum_{t=t_1}^{t_2} L(t)$, and $A_H(t_1, t_2) := \sum_{t=t_1}^{t_2} H(t)$. Note that $A_L(t_1, t_2)$ and $A_H(t_1, t_2)$ denote, respectively, the number of packets entering the light queue and the heavy queue in slots t_1 through t_2. For $y \in \mathbb{N}$, $A_L^{(y)}(t_1, t_2) := \sum_{t=t_1}^{t_2} L(t)\mathbf{1}_{\{L(t)\le y\}}$. Note that $A_L^{(y)}(t_1, t_2)$ is the number of packets entering the light queue from jobs of size $\le y$ in slots t_1 through t_2. Let $\lambda_L^{(y)} := \mathbb{E}[L(1)\mathbf{1}_{\{L(1)\le y\}}]$. It follows from the monotone convergence theorem that $\lim_{y\to\infty}\lambda_L^{(y)} = \lambda_L$. Let $\bar{p}_H := p_H(1 - p_L)$ denote the probability that a slot is exclusive for the heavy queue (recall that we have analogously defined $\bar{p}_L := p_L(1 - p_H)$). Finally, define $\bar{S}_L(t_1, t_2) := \sum_{t=t_1}^{t_2} \mathbf{1}_{\{\eta_L(t)=1,\eta_H(t)=0\}}$, and $\bar{S}_H(t_1, t_2) := \sum_{t=t_1}^{t_2} \mathbf{1}_{\{\eta_H(t)=1,\eta_L(t)=0\}}$. $\bar{S}_L(t_1, t_2)$ and $\bar{S}_H(t_1, t_2)$ denote, respectively, the number of exclusive slots available to the light queue and the heavy queue in slots t_1 through t_2. Note that in the wireline scenario, $\bar{S}_L(t_1, t_2) = \bar{S}_H(t_1, t_2) = 0$.

Next, we state the following lemma which is used repeatedly in our proofs; we give the proof in Appendix A.

Lemma 2: Let $\{A(t)\}$ denote a Bernoulli process taking values in $\{0, 1\}$, with $P(A(t) = 1) = p > 0$. Let B denote a random variable independent of $\{A(t)\}$ taking values in $\mathbb{N}$. Define

$$T := \min\left\{t \in \mathbb{N} \mid \sum_{i=1}^{t} A(i) \ge B\right\}.$$

The following statements are true.

(i) If B is light-tailed, then T is light-tailed.

(ii) If B is heavy-tailed with $\Gamma(B) \in (0, \infty)$, then $\Gamma(T) = \Gamma(B)$.

Finally, our results for the max-weight-α inter-queue scheduling rely on the following queue length tail asymptotics derived in [21, Chapter 5]. It is proved there that under max-weight-α inter-queue scheduling

$$\Gamma(q_L) = \alpha_L(\theta_H - 1), \qquad \Gamma(q_H) = \theta_H - 1. \tag{4}$$

We are now ready to prove the results claimed in the previous section. We prove these theorems in the order of their presentation in Section III.

A. Proof of Theorem 1

This section is devoted to the proof of Theorem 1. Our proof is based on formalizing the intuition that if a job of size $\Theta(x^{\alpha_L})$ arrives into the heavy queue early in the busy period, then with high probability, the light queue is denied service for a period of $\Omega(x)$ slots, except in its exclusive slots.

The proof relies on the following representation for the response-time tail. Consider a tagged busy period of the system. Let N_L denote the number of jobs entering the light queue in this busy period, and $V_{L,i}$, for $i = 1, 2, \cdots, N_L$, denote the response time of the i'th arriving job. The tail of V_L has the following well-known representation [27, Theorem 1.2, Ch. 6]

$$P(V_L > x) = \frac{\mathbb{E}\left[N_L^{(x)}\right]}{\mathbb{E}[N_L]} \tag{5}$$

where $N_L^{(x)} := \sum_{i=1}^{N_L} \mathbf{1}_{\{V_{L,i}>x\}}$ is the number of jobs in the light queue that experience a response time exceeding x in the busy period.[4] The proof proceeds by defining a 'bad' event $I(x)$ such that the bound

$$P(V_L > x) \ge \frac{P(I(x))\,\mathbb{E}\left[N_L^{(x)} \mid I(x)\right]}{\mathbb{E}[N_L]} \tag{6}$$

leads us to the statement of the theorem.

Without loss of generality, assume that the busy period under consideration starts in time slot 1. Recall that over the subset Λ' of the stability region, $\bar{p}_L < \lambda_L$, and $\lim_{y\to\infty}\lambda_L^{(y)} = \lambda_L$. Pick y large enough so that $\bar{p}_L < \lambda_L^{(y)}$. Let $\delta := (\lambda_L^{(y)} - \bar{p}_L)/4$.

We are now ready to define the event $I(x)$. Fix $\epsilon > 0$.

$$\begin{aligned}I(x) := {}&\left\{H(1) > \left\lceil\frac{xy}{\delta}\right\rceil + (\lambda_L+\epsilon)^{\alpha_L}\left\lceil\frac{xy}{\delta}\right\rceil^{\alpha_L}\right\}\bigcap\\ &\left\{A_L\left(1, \left\lceil\frac{xy}{\delta}\right\rceil\right) < (\lambda_L+\epsilon)\left\lceil\frac{xy}{\delta}\right\rceil\right\}\bigcap\\ &\left\{\bar{S}_L\left(1, \left\lceil\frac{xy}{\delta}\right\rceil\right) < (\bar{p}_L+\delta)\left\lceil\frac{xy}{\delta}\right\rceil\right\}\bigcap\\ &\left\{A_L^{(y)}\left(1, \left\lceil\frac{xy}{\delta}\right\rceil\right) > (\lambda_L^{(y)}-\delta)\left\lceil 1\frac{xy}{\delta}\right\rceil\right\}\\ &=: I_1(x) \cap I_2(x) \cap I_3(x) \cap I_4(x).\end{aligned}$$

Informally, the event $I_1(x)$ corresponds to the busy period starting with a 'large' job of size $O(x^{\alpha_L})$ entering the heavy queue. The events $I_2(x)$, $I_3(x)$, and $I_4(x)$ state that the number of packet arrivals into the light queue and number of exclusive slots for the light queue over the interval from slot 1 to slot $\lceil xy/\delta\rceil$ do not deviate much from their 'law of large numbers' estimates. Indeed, the weak law of large numbers implies that the events $I_2(x)$, $I_3(x)$, and $I_4(x)$ have a probability approaching 1 as $x \to \infty$.

Next, we show that the event $I(x)$ implies that at least x jobs entering the light queue in the busy period under consideration experience a response time exceeding x time slots. To see this, note that the event $I_1(x) \cap I_2(x)$ implies that the heavy queue has priority over the light queue in slots 1 through $\lceil xy/\delta\rceil$. Indeed, $I_1(x)$ implies that the length of the heavy queue remains greater than $(\lambda_L + \epsilon)^{\alpha_L}\lceil xy/\delta\rceil^{\alpha_L}$ over this interval, and $I_2(x)$ implies that the length of the light queue never exceeds $(\lambda_L + \epsilon)\lceil xy/\delta\rceil$ over the same interval. As a result, under event

[4] That $\mathbb{E}[N_L] < \infty$ may be justified as follows. It follows from the Lyapunov analysis in [7] that the tuple of queue occupancies evolves according to a positive recurrent Markov chain. This implies that busy periods of the queueing systems have finite mean, which in turn implies that $\mathbb{E}[N_L] < \infty$ via Wald's lemma.

$I(x)$, the light queue receives service only in its exclusive slots until time $\lceil xy/\delta \rceil$. Note that $I_3(x)$ gives an upper bound on the number of exclusive slots received by the light queue until time $\lceil xy/\delta \rceil$. Finally, $I_4(x)$ gives a lower bound on the number packets arriving into the light queue until time $\lceil xy/\delta \rceil$ from jobs of size $\leq y$. Therefore, under event $I(x)$, the number of packets remaining in the light queue after time slot $\lceil xy/\delta \rceil$, corresponding to jobs of size $\leq y$ exceeds

$$\left(\lambda_L^{(y)} - \delta\right) \left\lceil \frac{xy}{\delta} \right\rceil - (\bar{p}_L + \delta) \left\lceil \frac{xy}{\delta} \right\rceil = 2\delta \left\lceil \frac{xy}{\delta} \right\rceil \geq 2xy.$$

Now, since the corresponding jobs have a size of at most y, we conclude that under $I(x)$, the light queue contains at least $2x$ jobs at the end of $\lceil xy/\delta \rceil$ slots. Since each of these jobs requires at least one slot of service to complete, we conclude that under $I(x)$, at least $x-1$ jobs experience a response time exceeding x in the busy period under consideration.

Returning now to the bound (6), we have defined the event $I(x)$ such that $\mathbb{E}(N_L^{(x)) \mid I(x)} \geq x - 1$. To bound the probability of $I(x)$, note that

$$P\left(I(x)\right) = P\left(I_1(x)\right) P\left(I_2(x) \cap I_3(x) \cap I_4(x)\right)$$

since the arrival process into the heavy queue is independent of the arrival process into the light queue and the queue connectivity processes. Invoking the weak law of large numbers, we conclude that for $\nu \in (0,1)$, $P(I_2(x) \cap I_3(x) \cap I_4(x)) > (1-\nu)$ for large enough x. Therefore, for large enough x,

$$P(V_L > x) \geq \frac{1-\nu}{\mathbb{E}[N_L]}(x-1)\, P\left(I_1(x)\right).$$

The above statement implies that

$$\begin{aligned}\limsup_{x\to\infty} -\frac{\log PV_L > x}{\log(x)} &\leq \lim_{x\to\infty} -\frac{\log PI_1(x)}{\log(x)} - 1 \\ &= \alpha_L \theta_H - 1,\end{aligned}$$

where the last step above uses the fact that $H(1) \in \mathcal{RV}(\theta_H)$, which implies that

$$\lim_{x\to\infty} -\frac{\log P\left(H(1) > \left\lceil \frac{xy}{\delta} \right\rceil + (\lambda_L + \epsilon)^{\alpha_L} \left\lceil \frac{xy}{\delta} \right\rceil^{\alpha_L}\right)}{\log(x)} = \alpha_L \theta_H.$$

This completes the proof.

B. Proof of Theorem 2

This section is devoted to the proof of Theorem 2. The proof is relatively straightforward for the 'wireless' case, i.e., $p_L, p_H \in (0,1)$. We give the proof of this case here. The proof for the wireline scenario is more involved; we omit this proof here due to space constraints.[5] We refer the reader to [28, Chap. 5] for the proof.

Proof of Theorem 2 for the Wireless Scenario: Consider a tagged job entering the light queue in slot 0 in steady state. The tagged job has size $L(0) > 0$ and sees a queue length $q_L(0)$ in the light queue. Let us denote the response time of the tagged job by V_L. We need to prove that

$$\lim_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} = \alpha_L(\theta_H - 1). \tag{7}$$

We do this by proving matching asymptotic lower and upper bounds on the tail of V_L.

The lower bound on the tail of V_L is easy: since packets in the light queue are served in a FCFS manner, $V_L \geq q_L(0)$. Therefore, $P(V_L > x) \geq P(q_L(0) > x)$, which implies, using (4), that

$$\limsup_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} \leq \alpha_L(\theta_H - 1). \tag{8}$$

We now obtain the upper bound on the tail of V_L. Note that

$$q_L(1) = q_L(0) - \mathbf{1}_{\{\text{light queue for service in slot } 0\}} + L(0).$$

Since the light queue uses FCFS scheduling, V_L is simply equal to the time it takes for the light queue to receive service $q_L(1)$ times. Define $T := \min\{x \in \mathbb{N} \mid \bar{S}_L(1,x) \geq q_L(1)\}$. Note that T is the time it takes after slot 0 for the light queue to see $q_L(1)$ exclusive slots. Clearly, $V_L \leq T$. Since $q_L(0)$ is heavy-tailed with tail index $\alpha_L(\theta_H - 1)$, and $L(0)$ is light-tailed, it is easy to show that $q_L(1)$ is heavy-tailed with tail index $\alpha_L(\theta_H - 1)$. It follows then from Lemma 2 that T is also heavy-tailed with tail index $\alpha_L(\theta_H - 1)$. Since $V_L \leq T$, we obtain

$$\liminf_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} \geq \alpha_L(\theta_H - 1). \tag{9}$$

(8) and (9) of course imply (7). This completes the proof. □

C. Proof of Theorem 3

Informally, the proof of Theorem 3 is based on the following idea: a large arrival into the heavy queue early into the busy period can cause a large number of jobs in the light queue to experience a response time of the same order as the length of the busy period. We omit the proof here due to space constraints, noting that the proof is structurally similar to the proof of Theorem 7 in Section IV-G. We refer the reader to [28, Chap. 5] for the full proof.

D. Proof of Theorem 4

This section is devoted to the proof of Theorem 4. Consider a tagged job entering the heavy queue in slot 0 in steady state. The tagged job has size $H(0) > 0$. Let us denote the response time of the tagged job by V_H. We need to prove that

$$\lim_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} = \theta_H - 1.$$

We do this by proving matching asymptotic lower and upper bounds on the tail of V_H.

The lower bound is easy: since packets in the heavy queue are served in a FCFS manner, $V_H \geq q_H(0)$. Therefore, $P(V_H > x) \geq P(q_H(0) > x)$, which implies, using (4), that

$$\limsup_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} \leq \theta_H - 1.$$

[5] The analysis of the wireline case is more involved for the following reason. At its core, the proof requires a lower bound on the service process of the light queue. In the wireless case, the presence of exclusive slots provides a trivial lower bound on the service process of the light queue. However, in the wireline case, such a trivial bound is unavailable, and one needs to explicitly account for the stochastics of the two arrival processes in bounding the service process of the light queue.

We now obtain a matching upper bound. We do this separately for the wireless (i.e., $p_L, p_H \in (0,1)$) and the wireline case.

Wireless Case: Since the heavy queue uses FCFS scheduling, V_H is simply equal to the time it takes for the heavy queue to receive service $q_H(1)$ times. Define $T := \min\{x \in \mathbb{N} \mid \bar{S}_H(1,x) \geq q_H(1)\}$. Note that T is the time it takes after slot 0 for the heavy queue to see $q_H(1)$ exclusive slots. Clearly, $V_H \leq T$. Since $q_H(0)$ is heavy-tailed with tail index $\theta_H - 1$, and $H(0)$ is heavy-tailed with tail index θ_H, it is easy to show that $q_H(1)$ is heavy-tailed with tail index $\theta_H - 1$. It follows then from Lemma 2 that T is heavy-tailed with tail index $\theta_H - 1$. Since $V_H \leq T$,

$$\liminf_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} \geq \theta_H - 1.$$

This gives us the matching upper bound, and completes the proof for the wireless case.

Wireline Case: In the wireline case, we use the fact that $V_H \leq Z$, where Z is the number of time slots following the arrival of the tagged job till the system empties. We argue below that $Z \in \mathcal{RV}(\theta_H - 1)$. This implies that

$$\liminf_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} \geq \lim_{x\to\infty} -\frac{\log P(Z > x)}{\log(x)} = \theta_H - 1.$$

This gives us the required matching upper bound. It remains now to show that $Z \in \mathcal{RV}(\theta_H - 1)$. Note that in the wireline scenario, the sum of queue lengths evolves as a discrete time $Geo/GI/1$ queue in which the amount of work entering the queue in slot t equals $B(t) := L(t) + H(t)$. Z is simply the residual busy period for this $Geo/GI/1$ queue. Since $B(t)$ is regularly varying with index θ_H, it follows that the residual busy period is regularly varying with index $\theta_H - 1$.

E. *Proof of Theorem 5*

This section is devoted to the proof of Theorem 5. As in the previous proof, consider a tagged job entering the heavy queue in slot 0 in steady state. The tagged job has size $H(0) > 0$. Let us denote the response time of the tagged job by V_H. We need to prove that

$$\lim_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} = \theta_H.$$

We do this by proving matching asymptotic lower and upper bounds on the tail of V_H.

The lower bound is easy: it is clear that $V_H \geq H(0)$. Therefore,

$$\limsup_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} \leq \lim_{x\to\infty} -\frac{\log P\left(H(0) > x\right)}{\log(x)} = \theta_H.$$

To prove the upper bound, define

$$Z := \min\left\{x \in \mathbb{N} \mid \left(H(0) + q_L(1) + \sum_{i=1}^{x}(L(i) + H(i)) \leq x\right)\right\}.$$

Z is defined so that at the end of time slot Z, the total occupancy of both queues equals the number of packets waiting in the heavy queue at the time of the tagged job's arrival. Since the heavy queue uses PLCFS scheduling, it follows that $V_H \leq Z$.

We now invoke Lemma 8 in Appendix C to show that $Z \in \mathcal{RV}(\theta_H)$. Note that $H(0) \in \mathcal{RV}(\theta_H)$, and $q_H(1)$ is heavy-tailed with tail index $\alpha_L(\theta_H - 1)$. For the range of α_L under consideration, $\theta_H < \alpha_L(\theta_H - 1)$, i.e., $H(0)$ has a heavier tail than $q_H(1)$. This implies that $H(0) + q_H(1) \in \mathcal{RV}(\theta_H)$. Therefore, Lemma 8 in Appendix C implies that $Z \in \mathcal{RV}(\theta_H)$.

Finally, using the fact that $V_H \leq Z$, we have

$$\liminf_{x\to\infty} -\frac{\log P(V_H > x)}{\log(x)} \geq \lim_{x\to\infty} -\frac{\log P(Z > x)}{\log(x)} = \theta_H.$$

This gives us the desired matching upper bound, and completes the proof.

F. *Proof of Theorem 6*

This section is devoted to the proof of Theorem 6. We restrict ourselves to the 'wireless' case, i.e., $p_L, p_H \in (0,1)$, in this paper. The proof for the wireline case is considerably more involved, and can be found in [28, Chap. 5].[5] Our proof relies crucially on Theorem 8, which we prove first.

The proof of Theorem 8 utilizes a property (Lemma 3 below) of the class of long-tailed distributions, which is an important subclass of heavy-tailed distributions. Formally, a non-negative random variable X (or its d.f. F_X) is said to be *long-tailed* (denoted $X \in \mathcal{L}$) if $\lim_{x\to\infty}(P(X > x + y)/P(X > x)) = 1$ for all $y > 0$. The class of regularly varying distributions is a strict subset of the class of long-tailed distributions, which in turn is a strict subset of the class of heavy-tailed distributions [25]. We use the following sufficient condition for a distribution to be long-tailed (see Appendix B for the proof).

Lemma 3: Suppose X is a non-negative random variable. If $\Psi_X(\cdot)$ is eventually non-increasing, i.e., there exists $x_0 \geq 0$ such that $\Psi_X(\cdot)$ is non-increasing over $[x_0, \infty)$, and

$$\lim_{x\to\infty} \Psi_X(x) = 0,$$

then $X \in \mathcal{L}$.

Additionally, the proof of Theorem 8 relies on the following lemma.

Lemma 4: Suppose that F_X is the distribution function corresponding to a non-negative random variable X. If $X \in \mathcal{L}$, and $\bar{F}_X(x) := 1 - F_X(x)$ is strictly decreasing over $x \geq 0$, then under log-max-weight inter-queue scheduling,

$$\mathbb{E}\left[\frac{1}{\bar{F}_X(q_L)}\right] < \infty \; and \; \mathbb{E}\left[\frac{1}{\bar{F}_X\left(\log(1 + q_H)\right)}\right] < \infty.$$

The above lemma is a direct consequence of Theorem 1 in [7]. Note that if $X \in \mathcal{L}$, then $1/\bar{F}(x)$ grows sub-exponentially. Therefore, Lemma 4 states that certain *sub-exponential* moments of q_L are finite. However, in order to prove that q_L is light-tailed, we need to show that certain *exponential* moments of q_L are finite, i.e., $\mathbb{E}[e^{\beta q_L}] < \infty$ for some $\beta > 0$. We do this as follows.

Proof of Theorem 8: For the purpose of obtaining a contradiction, let us assume that q_L is heavy-tailed. Invoking Lemma 1, we conclude that $\liminf_{x\to\infty} \Psi_{q_L}(x) = 0$. Fix $\delta \in (0,1)$.

It is easy to argue that there exists a strictly increasing integer sequence $\{x_k\}_{k\geq 1}$, with $x_1 = 0$, and $x_k \overset{k\uparrow\infty}{\rightarrow} \infty$ such that

(i) $\{\Psi_{q_L}(x_k)\}_{k\geq 2}$ is non-increasing with $\lim_{k\to\infty}\Psi_{q_L}(x_k) = 0$,

(ii) $\bar{F}_{q_L}(x_{k+1}) \leq (1-\delta)\bar{F}_{q_L}(x_k)$ for $k \geq 1$.

We now define a distribution F_Y that agrees with F_{q_L} along the sequence $\{x_k\}$ such that F_Y satisfies the conditions of Lemma 4, implying that $\mathbb{E}[1/\bar{F}_Y(q_L)] < \infty$. We then show via a direct computation that $\mathbb{E}[1/\bar{F}_Y(q_L)] = \infty$. This gives us a contradiction, proving that q_L is light-tailed.

We define the distribution F_Y as follows. $\bar{F}_Y(x_k) = \bar{F}_{q_L}(x_k)$ for all $k \geq 1$. For $x \in (x_k, x_{k+1})$,

$$\log\left(\bar{F}_Y(x)\right) = \log\left(\bar{F}_Y(x_k)\right) + \frac{x - x_k}{x_{k+1}-x_k}\left(\log\left(\bar{F}_Y(x_{k+1})\right) - \log\left(\bar{F}_Y(x_k)\right)\right). \quad (10)$$

Note that for $x \in (x_k, x_{k+1})$, $\log(\bar{F}_Y(x))$ is defined by linearly interpolating between $\log(\bar{F}_Y(x_k))$ and $\log(\bar{F}_Y(x_{k+1}))$. Equation (10) implies, via simple algebraic manipulations that, for $x \in (x_k, x_{k+1})$, $k \geq 2$,

$$\begin{aligned}\Psi_Y(x) &= \frac{\log\left(\bar{F}_Y(x_k)\right) - \log\left(\bar{F}_Y(x_{k+1})\right)}{x_{k+1}-x_k} \\ &\quad + \frac{1}{x}\frac{x_k x_{k+1}\left(\Psi_Y(x_k) - \Psi_Y(x_{k+1})\right)}{x_{k+1}-x_k} \\ &=: \nu_{1,k} + \frac{\nu_{2,k}}{x},\end{aligned}$$

where $\nu_{1,k} > 0$, $\nu_{2,k} \geq 0$. Note that $\Psi_Y(x)$ is non-increasing over $x \in (x_k, x_{k+1})$ for $k \geq 2$. It then follows from Condition (i) above that $\Psi_Y(x)$ is non-increasing over $[x_2, \infty)$ with $\lim_{x\to\infty}\Psi_Y(x) = 0$. From Lemma 3, we conclude then that $Y \in \mathcal{L}$. Moreover, since $\bar{F}_Y(x)$ is strictly decreasing over $x \geq 0$ by definition, Lemma 4 implies that $\mathbb{E}[1/\bar{F}_Y(q_L)] < \infty$.

We now show through a direct computation that $\mathbb{E}[1/\bar{F}_Y(q_L)] = \infty$. Pick $k_0 \in \mathbb{N}$.

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\bar{F}_Y(q_L)}\right] &\geq \sum_{x=1}^{x_{k_0+1}} \frac{1}{\bar{F}_Y(x)}P(q_L = x) \\ &= \sum_{k=1}^{k_0}\sum_{x=x_k+1}^{x_{k+1}} \frac{1}{\bar{F}_Y(x)}P(q_L = x) \\ &\geq \sum_{k=1}^{k_0}\sum_{x=x_k+1}^{x_{k+1}} \frac{1}{\bar{F}_Y(x_k)}P(q_L = x) \\ &= \sum_{k=1}^{k_0}\frac{\bar{F}_{q_L}(x_k) - \bar{F}_{q_L}(x_{k+1})}{\bar{F}_Y(x_k)}.\end{aligned}$$

Now, since $\bar{F}_Y$ and $\bar{F}_{q_L}$ agree along the sequence $\{x_k\}$,

$$\mathbb{E}\left[\frac{1}{\bar{F}_Y(q_L)}\right] \geq \sum_{k=1}^{k_0}\frac{\bar{F}_Y(x_k) - \bar{F}_Y(x_{k+1})}{\bar{F}_Y(x_k)} \geq \sum_{k=1}^{k_0}\delta = k_0\delta, \quad (11)$$

where the last step above uses the fact that $\bar{F}_Y(x_{k+1}) \leq (1-\delta)\bar{F}_Y(x_k)$ for $k \geq 1$. Since $\mathbb{E}[1/\bar{F}_Y(q_L)] \geq k_0\delta$ for any $k_0 \in \mathbb{N}$, it follows that $\mathbb{E}[1/\bar{F}_Y(q_L)] = \infty$. This gives us a contradiction, which proves that q_L is light-tailed. ■

We are now ready to give the proof of Theorem 6 for the wireless case.

Proof of Theorem 6 for the Wireless Case: Consider a tagged job entering the light queue in slot 0 in steady state. The tagged job has size $L(0) > 0$ and sees a queue length $q_L(0)$ in the light queue. Theorem 8 implies that $q_L(0)$ is light-tailed.

Let us denote the response time of the tagged job by V_L. We need to prove that V_L is light-tailed. Note that

$$q_L(1) = q_L(0) - \mathbf{1}_{\{\text{light queue for service in slot } 0\}} + L(0).$$

Since $q_L(0)$ and $L(0)$ are both light-tailed, it follows that $q_L(1)$ is light-tailed. Now, since the light queue uses FCFS scheduling, V_L is simply equal to the time it takes for the light queue to receive service $q_L(1)$ times. Define $T := \min\{x \in \mathbb{N} \mid \bar{S}_L(1,x) \geq q_L(1)\}$. Note that T is the time it takes after slot 0 for the light queue to see $q_L(1)$ exclusive slots. Clearly, $V_L \leq T$. Now, it follows from Lemma 2 that T is light-tailed, which implies that V_L is light-tailed. This completes the proof of Theorem 6 for the wireless case.

G. Proof of Theorem 7

This section is devoted to the proof of Theorem 7. As in the proof of Theorem 1, our proof of Theorem 7 is based on defining a ‘bad’ event $I(x)$ in a tagged busy period, such that the bound (6) leads us to the statement of the theorem. Informally, the event $I(x)$ involves a large enough job arriving into the heavy queue to start the busy period, resulting in $\Omega(\log(x))$ jobs in the light queue experiencing a response time of $\Omega(x)$ slots in the busy period.

Without loss of generality, assume that the busy period under consideration starts in time slot 1. Recall that over the subset of interest Λ' of the stability region, $\bar{p}_L < \lambda_L$, and $\lim_{y\to\infty}\lambda_L^{(y)} = \lambda_L$. Pick y large enough so that $\bar{p}_L < \lambda_L^{(y)}$. Pick $\delta > 0$ such that $\delta \leq (\lambda_L^{(y)} - \bar{p}_L)/4$.

Our ‘bad’ event $I(x) := G(x) \cap J(x)$, where we define and interpret the events $G(x)$ and $J(x)$ below. We start with the definition of $G(x)$. This event is parameterized by $\beta \in \mathbb{N}$, whose value we fix later.

$$\begin{aligned}G(x) \coloneqq &\left\{H(1) > \beta\lfloor\log(x)\rfloor + x^{\beta(\lambda_L+2\delta)} + x + \beta\delta\log(x)\right\} \\ &\bigcap\left\{A_L\left(1, \beta\lfloor\log(x)\rfloor\right) < (\lambda_L+\delta)\beta\lfloor\log(x)\rfloor\right\} \\ &\bigcap\left\{\bar{S}_L\left(1, \beta\lfloor\log(x)\rfloor\right) < (\bar{p}_L+\delta)\beta\lfloor\log(x)\rfloor\right\} \\ &\bigcap\left\{A_L^{(y)}\left(1, \beta\lfloor\log(x)\rfloor\right) > (\lambda_L^{(y)}-\delta)\beta\lfloor\log(x)\rfloor\right\} \\ =:\ & G_1(x) \cap G_2(x) \cap G_3(x) \cap G_4(x).\end{aligned}$$

Roughly, $G(x)$ states that a job of size $\Theta(x^{\max\{\beta(\lambda_L+2\delta),1\}})$ arrives into the heavy queue at the start of the busy period, and the number of arrivals in the light queue, as well as the number of exclusive slots seen by it in slots 1 through $\beta\lfloor\log(x)\rfloor$ do not deviate much from their ‘law of large numbers’ estimates. The following lemma states a key implication of $G(x)$.

Lemma 5: $G(x)$ implies that at the end of $\beta\lfloor\log(x)\rfloor$ slots,

(i) the occupancy of the heavy queue strictly exceeds $x^{\beta(\lambda_L+2\delta)} + x + \beta\delta\log(x)$, i.e., $q_H(\beta\lfloor\log(x)\rfloor + 1) > x^{\beta(\lambda_L+2\delta)} + x + \beta\delta\log(x)$,

(ii) the occupancy of the light queue is strictly less that $(\lambda_L + \delta)\beta\lfloor\log(x)\rfloor$, i.e., $q_L(\beta\lceil\log(x)\rceil + 1) < (\lambda_L + \delta)\beta\lfloor\log(x)\rfloor$,

(iii) the light queue contains at least $2\beta\delta\lfloor\log(x)\rfloor$ packets from jobs of size $\leq y$.

Proof: The first two claims of the lemma are easy to verify. Indeed, Claim (i) is a consequence of event $G_1(x)$, and Claim (ii) is a consequence of event $G_2(x)$. We give the arguments for Claim (iii) below.

Note that $G(x)$ implies that the light queue does not receive service, except in its exclusive slots, in slots 1 through $\beta\lfloor\log(x)\rfloor$. Indeed, the event $G_2(x)$ guarantees that

$$q_L(t) < (\lambda_L + \delta)\beta\left\lfloor\log(x)\right\rfloor$$

during this period, whereas the event $G_1(x)$ implies that

$$\log\left(1 + q_H(t)\right) > (\lambda_L + 2\delta)\beta\log(x)$$

over the same period. Also, during the period from slot 1 to slot $\beta\lfloor\log(x)\rfloor$, $G_3(x)$ gives an upper bound on the number of exclusive slots received by the light queue, and $G_4(x)$ gives a lower bound on the number packets arriving into the light queue from jobs of size $\leq y$. Therefore, under event $G(x)$, the number of packets remaining in the light queue after time slot $\beta\lfloor\log(x)\rfloor$, corresponding to jobs of size $\leq y$ exceeds

$$(\lambda_L - \delta)\beta\left\lfloor\log(x)\right\rfloor - (\bar{p}_L + \delta)\beta\left\lfloor\log(x)\right\rfloor \geq 2\delta\beta\left\lfloor\log(x)\right\rfloor.$$

This verifies Claim (iii). ■

Now, invoking the weak law of large numbers, we know that $P(G_2(x) \cap G_3(x) \cap G_4(x))$ approaches 1 as $x \to \infty$. Therefore, fixing $\nu \in (0,1)$,

$$P\left(G(x)\right) \geq (1-\nu)P\left(G_1(x)\right) \text{ for large enough } x. \quad (12)$$

Next, we define the event $J(x)$. Let

$$n(x) := \left\lceil \frac{x}{\lfloor\beta\delta\left\lfloor\log(x)\right\rfloor\rfloor} \right\rceil, \quad m(x) := \lfloor\beta\delta\left\lfloor\log(x)\right\rfloor\rfloor.$$

The event $J(x)$ concerns arrivals into the light queue, and exclusive slots available to it over $n(x)m(x)$ slots following slot $\beta\lfloor\log(x)\rfloor$. Specifically, the event $J(x)$ states that the number of arrivals in the light queue, as well as the number of exclusive slots available to it, do not deviate much from the corresponding 'law of large numbers' estimates over $n(x)$ periods, each period being $m(x)$ slots long. For notational convenience, define $t[k] := \beta\lfloor\log(x)\rfloor + (k-1)m(x)$. Formally,

$$J(x) := \bigcap_{k=1,2,\cdots,n(x)} J_k(x),$$

where

$$\begin{aligned} J_k(x) := &\left\{\bar{S}_L\left(t[k]+1, t[k+1]\right) < (\bar{p}_L + \delta)m(x)\right\} \bigcap \\ &\left\{A_L\left(t[k]+1, t[k+1]\right) > (\lambda_L - \delta)m(x)\right\} \\ =:& J_{k,1}(x) \cap J_{k,2}(x). \end{aligned}$$

The following lemma states the key implication of our 'bad' event $I(x) = G(x) \cap J(x)$.

Lemma 6: The event $I(x) = G(x) \cap J(x)$ implies that for $k = 1, \cdots, n(x)$,

(i) $q_L(t[k]+1) \geq q_L(t[1]+1)$

(ii) For $i = 2, 3, \cdots, m(x)$,

$$q_L(t[k]+i) \geq q_L\left(t[1]+1\right) - m(x)$$

Proof: We first point out that over the $n(x)m(x)$ slots following slot $\beta\lfloor\log(x)\rfloor$, the event $G_1(x)$ implies that

$$\log\left(1 + q_H(t)\right) > (\lambda_L + 2\delta)\beta\log(x).$$

Therefore, if the light queue is to win service in a non-exclusive slot during this period, its occupancy must strictly exceed $(\lambda_L + 2\delta)\beta\log(x)$.

Note that Claim (i) above is trivially true for $k = 1$. We prove the lemma inductively as follows. We show that if Claim (i) is true for $k = k_0$, then Claim (ii) is true for $k = k_0$, and Claim (i) is true for $k = k_0 + 1$.

Accordingly, let us assume that Claim (i) is true for some $k = k_0$. That Claim (ii) holds for $k = k_0$ is obvious, since the occupancy of the light queue can decrease by at most $m(x)$ in $m(x)$ slots. To show that Claim (i) holds for $k = k_0 + 1$, we consider the following two cases.

Case 1: In slots $t[k_0]+1$ through $t[k_0+1]$, the light queue received service only in exclusive slots.

In this case, since $J_{k_0}(x)$ implies that the arrivals into the light queue outnumber the number of free slots over this period, it follows that $q_L(t[k_0+1]+1) \geq q_L(t[k_0]+1)$, which implies that Claim (i) holds for $k = k_0 + 1$ given that it holds for $k = k_0$.

Case 2: In slots $t[k_0]+1$ through $t[k_0+1]$, the light queue received service in a non-exclusive slot.

As we have argued before, this case implies that in some slot in the interval under consideration, the light queue occupancy strictly exceeded $(\lambda_L + 2\delta)\beta\log(x)$. Thus, at the end of this interval (of length $m(x)$ slots), the light queue occupancy must exceed

$$(\lambda_L + 2\delta)\beta\log(x) - m(x) \geq (\lambda_L + \delta)\beta\log(x).$$

Since we know from Lemma 5 that

$$q_L\left(t[1]+1\right) < (\lambda_L + \delta)\beta\left\lfloor\log(x)\right\rfloor,$$

it then follows that Claim (i) holds for $k = k_0 + 1$.

This completes the proof. ■

The above lemma states that under event $I(x)$, over $n(x)m(x)$ slots following slot $\beta\lfloor\log(x)\rfloor$, the occupancy of the light queue never dips more than $m(x)$ below its level after slot $\beta\lfloor\log(x)\rfloor$. Moreover, we know from Lemma 5 that under event $I(x)$, at the end of slot $\beta\lfloor\log(x)\rfloor$, there are at least $2\beta\delta\lfloor\log(x)\rfloor$ packets in the light queue from jobs of size $\leq y$. Therefore, since the light queue uses PLCFS, we conclude that at least $\beta\delta\lfloor\log(x)\rfloor$ packets, from jobs of size $\leq y$, stay in queue for more than $n(x)m(x)$ slots. Since $n(x)m(x) \geq x$, this in turn implies that under event $I(x)$, at least $\beta\delta\lfloor\log(x)\rfloor/y$

jobs in the light queue experience a response time exceeding x, i.e.,

$$\mathbb{E}\left[N_L^{(x)} \mid I(x)\right] \geq \frac{\beta\delta \lfloor \log(x) \rfloor}{y}. \tag{13}$$

Note that the Chernoff bound implies that there exists $\tau > 0$ such that $P(J_{k,1}(x)) \geq 1 - e^{-\tau m(x)}$ and $P(J_{k,2}(x)) \geq 1 - e^{-\tau m(x)}$. Therefore, $P(J_k(x)) \geq 1 - 2e^{-\tau m(x)}$, implying that

$$P\left(J(x)\right) \geq \left(1 - 2e^{-\tau m(x)}\right)^{n(x)}.$$

Let us fix $\beta > 1/\tau\delta$. For this choice of β, it is easy to show that $P(J(x)) \overset{x\uparrow\infty}{\to} 1$, implying that

$$P\left(J(x)\right) \geq (1-\nu) \text{ for large enough } x. \tag{14}$$

Returning to our bound (6), we now have, using (12), (13), and (14):

$$PV_L > x \geq \frac{(1-\nu)^2}{\mathbb{E}[N_L]} \frac{2\beta\delta \lfloor \log(x) \rfloor}{y} P\left(G_1(x)\right)$$

It then follows that

$$\begin{aligned}\limsup_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} &\leq \lim_{x\to\infty} -\frac{\log P\left(G_1(x)\right)}{\log(x)} \\ &= \theta_H \max\left\{\beta(\lambda_L + 2\delta), 1\right\}\end{aligned}$$

where the last step uses the fact that $H(1) \in \mathcal{RV}(\theta_H)$.

Since $\bar{\Gamma}(V_L) < \infty$, V_L is heavy-tailed. This completes the proof.

H. Proof of Theorem 9

The proof follows along similar lines as the corresponding proof for max-weight-α scheduling (i.e., the proof of Theorem 4 in Section IV-D), except that for the lower bound on the tail of V_H, we need to prove that $\bar{\Gamma}(q_H) \leq \theta_H - 1$. This is easy to show, since $q_H \geq_{\text{st}} \hat{q}_H$, where $\hat{q}_H$ is the stationary queue length of a $Geo/GI/1$ queue fed by the same arrival process as the heavy queue. Since $\Gamma(\hat{q}_H) = \theta_H - 1$, it follows that $\bar{\Gamma}(q_H) \leq \theta_H - 1$.

I. Proof of Theorem 10

The proof of Theorem 10 is similar to the proof of Theorem 5 in Section IV-E, and is omitted.

V. Concluding Remarks

In this paper, we consider a scenario in which a heavy-tailed (extremely variable) flow and a light-tailed (moderately variable) flow contend for service from a single server. Prior work [19]–[21] has focused on analysing the distribution of queue-lengths in this setting. It was proved in [19]–[21] that the classical max-weight inter-queue scheduling policy, while throughput optimal, leads to heavy-tailed queue-lengths for the light-tailed flow. Further, it was shown that generalized max-weight policies that give relative priority to the light-tailed flow can improve its queue-length tail, while maintaining throughput optimality. In particular, it was proved in [21] that under log-max-weight inter-queue scheduling, the queue-length corresponding to the light-tailed flow is light-tailed in the wireline setting.

In this work, we extend the above results, focusing on the distributions of the *response times* seen by the flows. Indeed, from the standpoint of applications sending/receiving information, response time is perhaps a more relevant metric than queue length. A key contribution of this work is that our analysis of response times brings into focus the impact of the intra-queue scheduling policies.[6] Indeed, our results reveal that *the response time distribution is highly sensitive to the intra-queue scheduling policy*, much more so than in the case of stochastically homogeneous flows. For example, note that under log-max-weight inter-queue scheduling, FCFS scheduling within the light queue results in light-tailed response times (Theorem 6), whereas PLCFS scheduling results in heavy-tailed response times (Theorem 7). Thus, our results highlight that in the presence of extreme stochastic variability between flows, a careful choice of inter-queue as well as intra-queue policies is crucial in order to achieve good response-time tail behavior.

Additionally, in the process of analysing response times, we also prove a key extension of the queue-length tail asymptotics in [21]. In particular, [21] proves that the queue-length corresponding to the light-tailed flow is light-tailed under log-max-weight in the wireline scenario. We prove, via a novel application of Lyapunov bounds from [7], that this is also true in the general wireless setting (Theorem 8).

The results in this paper motivate future research along several directions. An immediate goal would be to complete the analysis of the response-time tail of the heavy-tailed flow in our model, i.e., to extend Theorems 5, 9, and 10 to the general wireless model. We conjecture that the response-time tail index would remain unchanged in the wireless model.

Additionally, it would be interesting to generalize our simplified model to more realistic network settings. There are at-least two promising directions for generalization.

One potential model generalization is to consider a system of N parallel queues (where $N \geq 2$), connected to a single server via ON-OFF links. A subset of these queues would see a heavy-tailed arrival process. Interestingly, some of the results of this paper extend easily to this setting. Specifically, it is easy to prove the following results, using similar arguments to those used in this paper.

1) For any light-tailed flow i, there exists a subset Λ'_i of the capacity region over which max-weight scheduling results in heavy-tailed response times.
2) Under log-max-weight inter-queue scheduling, the stationary queue length distribution for each light-tailed flow is light-tailed (note that the proof of Theorem 8 does not depend on having only two queues).
3) Under log-max-weight inter-queue scheduling, and FCFS scheduling in all light queues, the response time distribution for all light queues is light-tailed, under the assumption that all links have a positive probability of being OFF.

Another potential model generalization is to consider a multi-hop wireless network with interference constraints, wherein a subset of the traffic flows are heavy-tailed. The

[6] Note that queue lengths are insensitive to the intra-queue scheduling policy.

first queue-length tail asymptotics in such a setting (under max-weigh-α inter-queue scheduling) have been derived recently in [29].

Appendix A
Proof of Lemma 2

In this section, we prove Lemma 2. Pick $\epsilon \in (0,p)$. We first obtain the following upper bound on the tail of T. For $t \in \mathbb{N}$,

$$\begin{aligned} P(T > t) &= P\left(\sum_{i=1}^{t} A(i) < B\right) \\ &= P\left(\sum_{i=1}^{t} A(i) < B;\ B \leq t(p-\epsilon)\right) \\ &\quad + P\left(\sum_{i=1}^{t} A(i) < B;\ B > t(p-\epsilon)\right) \\ &\leq P\left(\sum_{i=1}^{t} A(i) < t(p-\epsilon)\right) + P\left(B > t(p-\epsilon)\right). \end{aligned}$$

Bounding the first term above using the Chernoff bound, we conclude that there exists $\phi_1 > 0$ such that

$$P(T > t) \leq e^{-\phi_1 t} + P\left(B > t(p-\epsilon)\right). \tag{15}$$

Now, suppose that B is light-tailed. Then by definition, there exists $\phi_2 > 0$ such that $P(B > t) \leq e^{\phi_2 t}$ for large enough t. Therefore, from (15), we conclude that there exist $c, \phi > 0$ such that $P(T > t) \leq ce^{-\phi t}$ for large enough t, which implies that T is light-tailed. This completes the proof of Statement (i).

Next, suppose that B is heavy-tailed with $\Gamma(B) \in (0,\infty)$. It then follows easily from (15) that

$$\begin{aligned} \liminf_{t\to\infty} -\frac{\log P(T > t)}{\log(t)} &\geq \liminf_{t\to\infty} -\frac{\log P\left(B > t(p-\epsilon)\right)}{\log(t)} \\ &= \Gamma(B). \end{aligned}$$

Also, since $T \geq B$, it follows that $P(T > t) \geq P(B > t)$, which implies that

$$\limsup_{t\to\infty} -\frac{\log P(T > t)}{\log(t)} \leq \limsup_{t\to\infty} -\frac{\log P(B > t)}{\log(t)} = \Gamma(B).$$

It therefore follows that $\Gamma(T) = \Gamma(B)$.[7] This completes the proof of Statement (ii).

Appendix B
Proof of Lemma 3

In this section, we prove Lemma 3. Pick $y > 0$. For large enough x,

$$\begin{aligned} \frac{\bar{F}_X(x+y)}{\bar{F}_X(x)} &= \frac{e^{-(x+y)\Psi_X(x+y)}}{e^{-(x)\Psi_X(x)}} \\ &\geq e^{-[(x+y)\Psi_X(x) - x\Psi_X(x)]} \\ &= e^{-y\Psi_X(x)}. \end{aligned}$$

The second step above uses the fact that $\Psi_X(\cdot)$ is eventually non-increasing. Since $\Psi_X(x) \to 0$ as $x \to \infty$, the above bound implies that

$$\begin{aligned} \liminf_{x\to\infty} \frac{\bar{F}_X(x+y)}{\bar{F}_X(x)} &\geq \liminf_{x\to\infty} e^{-y\Psi_X(x)} \\ &= 1. \end{aligned}$$

Of course, since $\bar{F}_X(x+y)/\bar{F}(x) \leq 1$, it is obvious that $\limsup_{x\to\infty}(\bar{F}_X(x+y)/\bar{F}(x)) \leq 1$. It therefore follows that $\lim_{x\to\infty}(\bar{F}_X(x+y)/\bar{F}(x)) = 1$. This completes the proof.

Appendix C
Technical Lemmas

In this section, we state two technical lemmas that are used in our proofs.

The first concerns the probability of extremely large deviations of the running sum of regularly varying i.i.d. random variables from the mean.

Lemma 7: Suppose that $\{Y(i)\}_{i\geq 1}$ is an i.i.d. sequence of non-negative random variables with $Y(1) \in \mathcal{RV}(\beta)$, $\beta > 1$. Also, suppose $\{\psi(n)\}_{n\geq 1}$ is a positive, increasing sequence that is superlinear, i.e., $\lim_{n\to\infty}(\psi(n)/n) = \infty$. Then

$$P\left(\sum_{i=1}^{n} Y(i) > \psi(n)\right) \sim nP\left(Y(1) > \psi(n)\right).$$

Lemma 8: Intuitively, the above lemma means that extremely large deviations of the running sum from its expected value occur most likely because of one extremely large value. The proof is quite involved, but it omitted here as it follows along the same lines as the proof of Example 23 in [30].

Lemma 9: Suppose $\{Y_i\}_{i\geq 1}$ is an i.i.d. sequence of non-negative random variables taking values in $\{0\} \cup \mathbb{N}$ satisfying $Y_i \in \mathcal{RV}(\beta)$ for $\beta > 1$, and $\mathbb{E}(Y_i) < 1$. Also, suppose that X is a non-negative random variable independent of $\{Y_i\}_{i\geq 1}$, such that X takes values in $\mathbb{N}$, and $X \in \mathcal{RV}(\theta)$, where $\theta > 0$. Define

$$T := \left\{t \in \mathbb{N} \mid X + \sum_{i=1}^{t} Y_i \leq t\right\}.$$

If $\theta \leq \beta$, then $T \in \mathcal{RV}(\theta)$.

In the above lemma, T may be interpreted as a busy period in a discrete-time $Geo/GI/1$ queue, started by a job of size X in the queue in time slot 0, with Y_i denoting the amount of work entering the queue in time slot i. The lemma states that if the busy period is started by a random variable X with tail heavier than the job size distribution, then the residual busy period has the same index as X. The proof is a straightforward application of Tauberian theorems that relate the tail of a regularly varying distribution and the behavior of its Laplace-Stieltjes transform around the origin [31] (see also Section III in [32]). We omit the proof here.

[7] Even though we have taken the limit of $-\log P(T > t)/\log(t)$ as $t \to \infty$ over $\mathbb{N}$, it is easy to show that the same limit holds as $t \to \infty$ over $\mathbb{R}$.

References

[1] J. Nair, K. Jagannathan, and A. Wierman, "When heavy-tailed and lighttailed flows compete: The response time tail under generalized maxweight scheduling," in *Proc. IEEE INFOCOM*, 2013.

[2] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 466–478, 1993.

[3] L. Tassiulas and A. Ephremides, “Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks,” *IEEE Trans. Autom. Contr.*, vol. 37, no. 12, pp. 1936–1948, 1992.

[4] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, “Achieving 100% throughput in an input-queued switch,” *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1260–1267, 1999.

[5] M. Neely, E. Modiano, and C. Rohrs, “Power and server allocation in a multi-beam satellite with time varying channels,” in *Proc. IEEE INFOCOM*, 2002.

[6] A. Stolyar, “Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic,” *Annals Appl. Probab.*, vol. 14, no. 1, pp. 1–53, 2004.

[7] A. Eryilmaz, R. Srikant, and J. R. Perkins, “Stable scheduling policies for fading wireless channels,” *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 411–424, 2005.

[8] A. Brzezinski and E. Modiano, “Dynamic reconfiguration and routing algorithms for IP-over-WDM networks with stochastic traffic,” *J. Lightw. Technol.*, vol. 23, no. 10, pp. 3188–3205, 2005.

[9] M. Neely, “Delay analysis for max weight opportunistic scheduling in wireless systems,” *IEEE Trans. Autom. Contr.*, vol. 54, no. 9, pp. 2137–2150, 2009.

[10] M. Neely, E. Modiano, and C. Rohrs, “Dynamic power allocation and routing for time varying wireless networks,” in *Proc. IEEE INFOCOM*, 2003.

[11] L. Le, K. Jagannathan, and E. Modiano, “Delay analysis of maximum weight scheduling in wireless *ad hoc* networks,” in *Proc. IEEE CISS*, 2009.

[12] A. Eryilmaz and R. Srikant, “Asymptotically tight steady-state queue length bounds implied by drift conditions,” *Queueing Syst.*, vol. 72, no. 3—4, pp. 311–359, 2012.

[13] A. Ganti, E. Modiano, and J. Tsitsiklis, “Optimal transmission scheduling in symmetric communication models with intermittent connectivity,” *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 998–1008, 2007.

[14] M. Neely, “Delay analysis for max weight opportunistic scheduling in wireless systems,” *IEEE Trans. Autom. Contr.*, vol. 54, no. 9, pp. 2137–2150, 2009.

[15] 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, 2003.

[16] S. Borst, O. Boxma, and M. Van Uitert, “The asymptotic workload behavior of two coupled queues,” *Queueing Syst.*, vol. 43, no. 1, pp. 81–102, 2003.

[17] O. Boxma, Q. Deng, and A. Zwart, “Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers,” *Queueing Syst.*, vol. 40, no. 1, pp. 5–31, 2002.

[18] S. Borst, R. Núñez-Queija, and B. Zwart, “Bandwidth sharing with heterogeneous flow sizes,” *Annals Telecommun.*, vol. 59, no. 11, pp. 1300–1314, 2004.

[19] M. Markakis, E. Modiano, and J. Tsitsiklis, “Scheduling policies for single-hop networks with heavy-tailed traffic,” in *Proc. 47th Annual Allerton Conf. Communication, Control, Computing*, 2009.

[20] K. Jagannathan, M. Markakis, E. Modiano, and J. Tsitsiklis, “Throughput optimal scheduling in the presence of heavy-tailed traffic,” in *Proc. 48th Annual Allerton Conf. Communication, Control, Computing*, 2010.

[21] K. Jagannathan, “Asymptotic Performance of Queue Length Based Network Control Policies,” Ph.D. dissertation, MIT, Cambridge, MA, USA, 2010.

[22] L. Huang, S. Moeller, M. Neely, and B. Krishnamachari, “Lifoback-pressure achieves near optimal utility-delay tradeoff,” in *Proc. WiOpt*, 2011.

[23] O. Boxma and B. Zwart, “Tails in scheduling,” *Perf. Eval. Rev.*, vol. 34, no. 4, pp. 13–20, 2007.

[24] S. Foss, D. Korshunov, and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions*. New York, NY, USA: Springer, 2011.

[25] K. Sigman, “Appendix: A primer on heavy-tailed distributions,” *Queueing Syst.*, vol. 33, no. 1, pp. 261–275, 1999.

[26] S. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York, NY, USA: Springer, 2007.

[27] S. Asmussen, *Applied Probability and Queues*. New York, NY, USA: Springer, 2003.

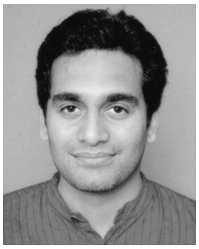
[28] J. Nair, “Scheduling for Heavy-Tailed and Light-Tailed Workloads in Queueing Systems,” Ph.D. dissertation, California Institute of Technology, Pasadena, CA, USA, 2012.

[29] 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, 2014.

[30] G. Samorodnitsky, Long Range Dependence, Heavy Tails and Rare Events 2002, Lecture notes [Online]. Available: http://dspace.library.cornell.edu/bitstream/1813/9228/1/TR001350.pdf

[31] N. N. H. Bingham and R. A. Doney, “Asymptotic properties of supercritical branching processes I: The Galton-Watson process,” in *Adv. Appl. Probab.*, 1974, pp. 711–731.

[32] S. Borst, O. Boxma, R. Núñez-Queija, and B. Zwart, “The impact of the service discipline on delay asymptotics,” *Perf. Eval.*, vol. 54, pp. 175–206, 2003.



Jayakrishnan Nair received the B.Tech. and M.Tech. in electrical engineering (EE) from IIT Bombay, India, 2007, and the Ph.D. degree in EE from California Institute of Technology, Pasadena, CA, USA, in 2012.

He has held post-doctoral positions at California Institute of Technology and Centrum Wiskunde & Informatica. He is currently an Assistant Professor in EE at IIT Bombay, India. His research focuses on modeling, performance evaluation, and design issues in queueing systems and communication networks.

Krishna Jagannathan received the B.Tech. degree in electrical engineering from IIT Madras, India, in 2004, and the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2006 and 2010, respectively.

During 2010–2011, he was a visiting post-doctoral scholar in Computing and Mathematical Sciences at Caltech, and an off-campus post-doctoral fellow at MIT. Since November 2011, he has been an Assistant Professor in the Department of Electrical Engineering, IIT Madras. He worked as a consultant at the Mathematical Sciences Research Center, Bell Laboratories, Murray Hill, NJ, USA, in 2005, and as an engineering intern at Qualcomm, Campbell, CA in 2007. His research interests lie in the stochastic modeling and analysis of communication networks, transportation networks, network control, and queuing theory. He is a member of IEEE and ACM.

Adam Wierman is a Professor in the Department of Computing and Mathematical Sciences at the California Institute of Technology, Pasadena, CA, USA, where he is a founding member of the Rigorous Systems Research Group (RSRG) and maintains a popular blog called Rigor+Relevance.

His research interests center around resource allocation and scheduling decisions in computer systems and services. He received the 2011 ACM SIGMETRICS Rising Star award, the 2014 IEEE Communications Society William R. Bennett Prize, and has been coauthor on papers that received best paper awards at ACM SIGMETRICS, IEEE INFOCOM, IFIP Performance (twice), IEEE Green Computing Conference, IEEE Power & Energy Society General Meeting, and ACM GREENMETRICS.