OPTIMAL QUEUE-SIZE SCALING IN SWITCHED NETWORKS

BY D. SHAH[1], N. S. WALTON AND Y. ZHONG[1]

*Massachusetts Institute of Technology, University of Amsterdam
and Columbia University*

We consider a switched (queuing) network in which there are constraints on which queues may be served simultaneously; such networks have been used to effectively model input-queued switches and wireless networks. The scheduling policy for such a network specifies which queues to serve at any point in time, based on the current state or past history of the system. In the main result of this paper, we provide a new class of online scheduling policies that achieve optimal queue-size scaling for a class of switched networks including input-queued switches. In particular, it establishes the validity of a conjecture (documented in Shah, Tsitsiklis and Zhong [*Queueing Syst.* **68** (2011) 375–384]) about optimal queue-size scaling for input-queued switches.

1. Introduction. A switched network consists of a collection of, say, N queues, operating in discrete time. At each time slot, queues are offered service according to a *service schedule* chosen from a specified finite set, denoted by $\mathcal{S}$. The rule for choosing a schedule from $\mathcal{S}$ at each time slot is called the *scheduling policy*. New work may arrive to each queue at each time slot exogenously and work served from a queue may join another queue or leave the network. We shall restrict our attention, however, to the case where work arrives in the form of unit-sized packets, and once it is served from a queue, it leaves the network, that is, the network is single-hop.

Switched networks are special cases of what Harrison [15, 16] calls "stochastic processing networks." Switched networks are general enough to model a variety of interesting applications. For example, they have been used to effectively model input-queued switches, the devices at the heart of high-end Internet routers, whose underlying silicon architecture imposes constraints on which traffic streams can be transmitted simultaneously [8]. They have also been used to model multihop wireless networks in which interference limits the amount of service that can be given

to each host [35]. Finally, they can be instrumental in finding the right operational point in a data center [31].

In this paper, we consider *online* scheduling policies, that is, policies that only utilize historical information (i.e., past arrivals and scheduling decisions). The performance objective of interest is the total queue size or total number of packets waiting to be served in the network on average (appropriately defined). The questions that we wish to answer are: (a) what is the minimal value of the performance objective among the class of online scheduling policies, and (b) how does it depend on the network structure, $\mathcal{S}$, as well as the effective load.

Consider a work-conserving $M/D/1$ queue with a unit-rate server in which unit-sized packets arrive as a Poisson process with rate $\rho \in (0, 1)$. Then, the long-run average queue-size scales[2] as $1/(1 - \rho)$. Such scaling dependence of the average queue size on $1/(1 - \rho)$ (or the inverse of the *gap*, $1 - \rho$, from the load to the capacity) is a universally observed behavior in a large class of queuing networks. In a switched network, the scaling of the average total queue size ought to depend on the number of queues, N. For example, consider N parallel $M/D/1$ queues as described above. Clearly, the average total queue size will scale as $N/(1 - \rho)$. On the other hand, consider a variation where all of these queues pool their resources into a single server that works N times faster. Equivalently, by a time change, let each of the N queues receive packets as an independent Poisson process of rate ρ/N, and each time a common unit-rate server serves a packet from one of the nonempty queues. Then, the average total queue-size scales as $1/(1 - \rho)$. Indeed, these are instances of switched networks that differ in their scheduling set $\mathcal{S}$, which leads to different queue-size scalings. Therefore, a natural question is the determination of queue-size scaling in terms of $\mathcal{S}$ and $(1 - \rho)$, where ρ is the effective load. In the context of an n-port input-queued switch with $N = n^2$ queues, the optimal scaling of average total queue size has been conjectured to be $n/(1 - \rho)$, that is, $\sqrt{N}/(1 - \rho)$ [29].

As the main result of this paper, we propose a new online scheduling policy for any single-hop switched network. This policy effectively emulates an insensitive bandwidth sharing network with a product-form stationary distribution with each component of this product-form behaving like an $M/M/1$ queue. This crisp description of stationary distribution allows us to obtain precise bounds on the average queue sizes under this policy. This leads to establishing, as a corollary of our result, the validity of a conjecture stated in [29] for input-queued switches. In general, it provides explicit bounds on the average total queue size for any switched network. Furthermore, due to the explicit bound on the stationary distribution of queue sizes under our policy, we are able to establish a form of large-deviations optimality of the policy for a large class of single-hop switched networks, including

---

[2]In this paper, by scaling of quantity we mean its dependence (ignoring universal constants) on $\frac{1}{1-\rho}$ and/or the number of queues, N, as these quantities become large. Of particular interest is the scaling of $\rho \to 1$ and $N \to \infty$, in that order.

the input-queued switches, and the independent-set model of wireless networks, when the underlying interference graph is bipartite, for example, and more generally, perfect.

The conjecture from [29] that we settle in this paper, states that in the heavy-traffic regime (i.e., $\rho \to 1$), the optimal average total queue-size scales as $\sqrt{N}/(1 - \rho)$. The validity of this conjecture is a significant improvement over the best-known bounds of $O(N/(1 - \rho))$ (due to the moment bounds of [24] for the maximum weight policy) or $O(\sqrt{N} \log N/(1 - \rho)^2)$ (obtained by using a batching policy [25]).

Our analysis consists of two principal components. First, we propose and analyze a scheduling mechanism that is able to emulate, in discrete time, *any* continuous-time bandwidth allocation within a bounded degree of error. This scheduler maintains a continuous-time queuing process and tracks its own queue size process. If, valued under a certain decomposition, the gap between the idealized continuous-time process and the real queuing process becomes too large, then an appropriate schedule is allocated. Second, we implement specific bandwidth allocation named the store-and-forward allocation policy (SFA). This policy was first considered by Massoulié, and was consequently discussed in the thesis of Proutière [26], Section 3.4. It was shown to be insensitive with respect to phase-type service distributions in works by Bonald and Proutière [3, 4]. The insensitivity of this policy for general service distributions was established by Zachary [41]. The store-and-forward policy is closely related to the classical product-form multi-class queuing network, which have highly desirable queue-size scalings. By emulating these queuing networks, we are able to translate results which render optimal queue-size bounds for a switched network. An interested reader is referred to [38] and [20] for an in-depth discussion on the relation between this policy, the proportionally fair allocation, and multi-class queuing networks.

1.1. *Organization.* In Section 2, we specify a stochastic switched network model. In Section 3, we discuss related works. Section 4 details the necessary background on the insensitive store-and-forward bandwidth allocation (SFA) policy. The main result of the paper is presented and proved in Section 5. We first describe the policy for single-hop switched networks, and state our main result, Theorem 5.2. This is followed by a discussion of the optimality of the policy. We then provide a proof of Theorem 5.2. A discussion of directions for future work is provided in Section 6.

Notation. Let $\mathbb{N}$ be the set of natural numbers $\{1, 2, \ldots\}$, let $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, let $\mathbb{R}$ be the set of real numbers and let $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$. Let $\mathbb{I}[A]$ be the indicator function of an event A, Let $x \wedge y = \min(x, y)$, $x \vee y = \max(x, y)$ and $[x]^+ = x \vee 0$. When $\mathbf{x}$ is a vector, the maximum is taken componentwise.

We will reserve bold letters for vectors in $\mathbb{R}^N$, where N is the number of queues. For example, $\mathbf{x} = [x_n]_{1 \leq n \leq N}$. Superscripts on vectors are used to denote labels, not

exponents, except where otherwise noted; thus, for example, $(\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2)$ refers to three arbitrary vectors. Let $\mathbf{0}$ be the vector of all 0s and $\mathbf{1}$ the vector of all 1s. The vector $\mathbf{e}_i$ is the ith unit vector, with all components being 0 but the ith component equal to 1. We use the norm $|\mathbf{x}| = \max_n |x_n|$. For vectors $\mathbf{u}$ and $\mathbf{v}$, we let $\mathbf{u} \cdot \mathbf{v} = \sum_{n=1}^N u_n v_n$. Let $\mathbf{A}^\mathsf{T}$ be the transpose of matrix $\mathbf{A}$. For a set $\mathcal{S} \subset \mathbb{R}^N$, denote its convex hull by $\langle \mathcal{S} \rangle$. For $n \in \mathbb{N}$, let $n! = \prod_{\ell=1}^n \ell$ be the factorial of n, and by convention, $0! = 1$.

2. Switched network model. We now introduce the switched network model. Section 2.1 describes the general system model, Section 2.2 lists the probabilistic assumptions about the arrival process and Section 2.3 introduces some useful definitions.

2.1. *Queueing dynamics.* Consider a collection of N queues. Let time be discrete, and indexed by $\tau \in \{0, 1, \ldots\}$. Let $Q_i(\tau)$ be the amount of work in queue $i \in \{1, \ldots, N\}$ at time slot τ. Following our general notation for vectors, we write $\mathbf{Q}(\tau)$ for $[Q_i(\tau)]_{1 \le i \le N}$. The initial queue sizes are $\mathbf{Q}(0)$. Let $A_i(\tau)$ be the total amount of work arriving to queue i, and $B_i(\tau)$ be the cumulative potential service to queue n, up to time τ, with $\mathbf{A}(0) = \mathbf{B}(0) = \mathbf{0}$.

We first define the queuing dynamics for a single-hop switched network. Defining $d\mathbf{A}(\tau) = \mathbf{A}(\tau + 1) - \mathbf{A}(\tau)$ and $d\mathbf{B}(\tau) = \mathbf{B}(\tau + 1) - \mathbf{B}(\tau)$, the basic Lindley recursion that we will consider is

$$(1) \qquad \mathbf{Q}(\tau + 1) = \left[\mathbf{Q}(\tau) - d\mathbf{B}(\tau)\right]^+ + d\mathbf{A}(\tau),$$

where the operation $[\cdot]^+$ is applied componentwise. The fundamental switched network constraint is that there is some finite set $\mathcal{S} \subset \mathbb{R}_+^N$ such that

$$(2) \qquad d\mathbf{B}(\tau) \in \mathcal{S} \qquad \text{for all } \tau.$$

For the purpose of this work, we shall focus on $\mathcal{S} \subset \{0, 1\}^N$. We will refer to $\boldsymbol{\sigma} \in \mathcal{S}$ as a schedule and $\mathcal{S}$ as the set of allowed schedules. In the applications in this paper, the schedule is chosen based on current queue sizes, which is why it is natural to write the basic Lindley recursion as (1) rather than the more standard $[\mathbf{Q}(\tau) + d\mathbf{A}(\tau) - d\mathbf{B}(\tau)]^+$.

For the analysis in this paper, it is useful to keep track of two other quantities. Let $Z_i(\tau)$ be the cumulative amount of idling at queue n, defined by $\mathbf{Z}(0) = \mathbf{0}$ and

$$(3) \qquad d\mathbf{Z}(\tau) = \left[d\mathbf{B}(\tau) - \mathbf{Q}(\tau)\right]^+,$$

where $d\mathbf{Z}(\tau) = \mathbf{Z}(\tau + 1) - \mathbf{Z}(\tau)$. Then, (1) can be rewritten as

$$(4) \qquad \mathbf{Q}(\tau) = \mathbf{Q}(0) + \mathbf{A}(\tau) - \mathbf{B}(\tau) + \mathbf{Z}(\tau).$$

Also, let $S_{\boldsymbol{\sigma}}(\tau)$ be the cumulative amount of time that is spent on using schedule $\boldsymbol{\sigma}$ up to time τ, so that

$$(5) \qquad \mathbf{B}(\tau) = \sum_{\boldsymbol{\sigma} \in \mathcal{S}} S_{\boldsymbol{\sigma}}(\tau) \boldsymbol{\sigma}.$$

A policy that decides which schedule to choose at each time slot $\tau \in \mathbb{Z}_+$ is called a *scheduling policy*. In this paper, we will be interested in online scheduling policies. That is, the scheduling decision at time τ will be based on historical information, that is, the cumulative arrival process $\mathbf{A}(\cdot)$ till time τ.

2.2. *Stochastic model.*   We shall assume that the exogenous arrival process for each queue is independent and Poisson. Specifically, unit-sized packets arrive to queue i as a Poisson process of rate λ_i. Let $\boldsymbol{\lambda} = [\lambda_i]_{i=1}^N$ denote the vector of all arrival rates. The results presented in this paper extend to more general arrival process with i.i.d. interarrival times with finite means, using a *Poissonization* trick. We discuss this extension in Section 6.

2.3. *Useful quantities.*   We shall assume that the scheduling constraint set $\mathcal{S}$ is *monotone*. This is captured in the following assumption.

ASSUMPTION 2.1 (Monotonicity).   If $\mathcal{S}$ contains a schedule, then $\mathcal{S}$ also contains all of its sub-schedules. Formally, for any $\boldsymbol{\sigma} \in \mathcal{S}$, if $\boldsymbol{\sigma}' \in \{0, 1\}^N$ and $\boldsymbol{\sigma}' \leq \boldsymbol{\sigma}$ componentwise, then $\boldsymbol{\sigma}' \in \mathcal{S}$.

Without loss of generality, we will assume that each unit vector $\mathbf{e}_i$ belongs to $\mathcal{S}$. Next, we define some quantities that will be useful in the remainder of the paper.

DEFINITION 2.2 (Admissible region).   Let $\mathcal{S} \subset \{0, 1\}^N$ be the set of allowed schedules. Let $\langle \mathcal{S} \rangle$ be the convex hull of $\mathcal{S}$, that is,

$$\langle \mathcal{S} \rangle = \left\{ \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma} : \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} = 1 \text{ and } \alpha_{\boldsymbol{\sigma}} \geq 0, \text{ for all } \boldsymbol{\sigma} \right\}.$$

Define the *admissible region* $\mathcal{C}$ to be

$$\mathcal{C} = \left\{ \boldsymbol{\lambda} \in \mathbb{R}_+^N : \boldsymbol{\lambda} \leq \boldsymbol{\sigma} \text{ componentwise, for some } \boldsymbol{\sigma} \in \langle \mathcal{S} \rangle \right\}.$$

Note that under Assumption 2.1, the capacity region $\mathcal{C}$ and the convex hull $\langle \mathcal{S} \rangle$ of $\mathcal{S}$ coincide.

Given that $\langle \mathcal{S} \rangle$ is a polytope contained in $[0, 1]^N$, there exists an integer $J \geq 1$, a matrix $\mathbf{R} \in \mathbb{R}_+^{J \times N}$ and a vector $\mathbf{C} \in \mathbb{R}_+^J$ such that

(6) $$\langle \mathcal{S} \rangle = \left\{ \mathbf{x} \in [0, 1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C} \right\}.$$

We call J the *rank* of $\langle \mathcal{S} \rangle$ in the representation (6). When it is clear from the context, we simply call J the rank of $\langle \mathcal{S} \rangle$. Note that this rank may be different from the rank of matrix $\mathbf{R}$. Our results will exploit the fact that the rank J may be an order of magnitude smaller than N.

DEFINITION 2.3 (Static planning problems and load).   Define the static planning optimization problem PRIMAL($\boldsymbol{\lambda}$) for $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ to be

$$\text{(7)} \qquad \text{minimize} \qquad \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}},$$

$$\text{(8)} \qquad \text{subject to} \qquad \boldsymbol{\lambda} \leq \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma},$$

$$\text{(9)} \qquad\qquad\qquad \alpha_{\boldsymbol{\sigma}} \in \mathbb{R}_+ \qquad \text{for all } \boldsymbol{\sigma} \in \mathcal{S}.$$

Define the induced load by $\boldsymbol{\lambda}$, denoted by $\rho(\boldsymbol{\lambda})$, as the value of the optimization problem PRIMAL($\boldsymbol{\lambda}$).

Note that $\boldsymbol{\lambda}$ is admissible if and only if $\rho(\boldsymbol{\lambda}) \leq 1$. It also follows immediately from Definition 2.3 that

$$\text{(10)} \qquad\qquad \rho(\boldsymbol{\lambda}) = \inf\{\gamma \geq 0 : \mathbf{R}\boldsymbol{\lambda} \leq \gamma \mathbf{C}\},$$

and $\boldsymbol{\lambda}$ is admissible if and only if $\mathbf{R}\boldsymbol{\lambda} \leq \mathbf{C}$, componentwise.

In the sequel, we will often consider the quantities $\tilde{\rho}_j = \sum_i R_{ji}\lambda_i / C_j$, for $j \in \{1, 2, \ldots, J\}$, which can be interpreted as loads on individual "resources" of the system (this interpretation will be made precise in Section 4). They are closely related to the system load $\rho(\boldsymbol{\lambda})$. We formalize this relation in the following lemma, whose proof is straightforward and omitted.

LEMMA 2.4.   *Consider a nonnegative matrix* $\mathbf{R} \in \mathbb{R}_+^{J \times N}$ *and a vector* $\mathbf{C} \in \mathbb{R}^J$ *with* $C_j > 0$ *for all* j. *For a nonnegative vector* $\boldsymbol{\lambda} \in \mathbb{R}_+^N$, *define* $\rho(\boldsymbol{\lambda})$ *by* (10) *and* $\tilde{\rho}_j = (\sum_i R_{ji}\lambda_i)/C_j$. *Then* $\rho(\boldsymbol{\lambda}) = \max_j \tilde{\rho}_j$.

The following is a simple and useful property of $\rho(\cdot)$: for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^N$,

$$\text{(11)} \qquad\qquad \rho(\mathbf{a} + \mathbf{b}) \leq \rho(\mathbf{a}) + \rho(\mathbf{b}).$$

2.4. *Motivating example.*   An Internet router has several input ports and output ports. A data transmission cable is attached to each of these ports. Packets arrive at the input ports. The function of the router is to work out which output port each packet should go to, and to transfer packets to the correct output ports. This last function is called *switching*. There are a number of possible switch architectures; we will consider the commercially popular input-queued switch architecture.

Figure 1 illustrates an input-queued switch with three input ports and three output ports. Packets arriving at input k destined for output ℓ are stored at input port k, in queue $Q_{k,\ell}$, thus there are $N = 9$ queues in total. (For this example, it is more natural to use double indexing, e.g., $Q_{3,2}$, whereas for general switched networks it is more natural to use single indexing, e.g., Q_i for $1 \leq i \leq N$.)

The switch operates in discrete time. At each time slot, the switch fabric can transmit a number of packets from input ports to output ports, subject to the two
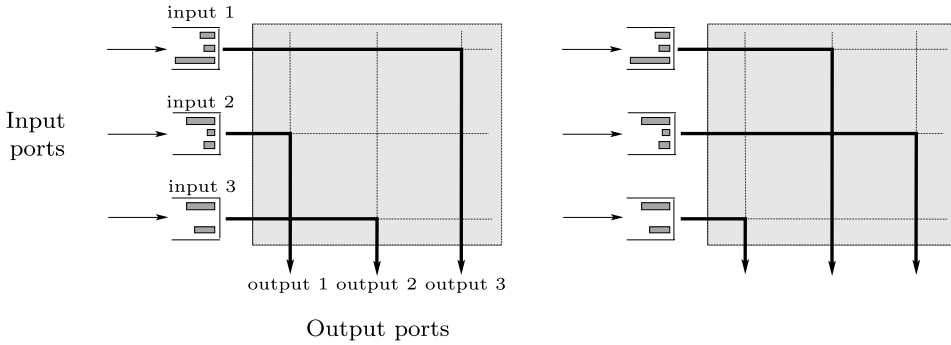
FIG. 1. *An input-queued switch, and two example matchings of inputs to outputs.*

constraints that each input can transmit at most one packet, and that each output can receive at most one packet. In other words, at each time slot the switch can choose a *matching* from inputs to outputs. The schedule $\sigma \in \mathbb{R}_+^{3\times 3}$ is given by $\sigma_{k,\ell} = 1$ if input port k is matched to output port ℓ in a given time slot, and $\sigma_{k,\ell} = 0$ otherwise. The matching constraints require that $\sum_{m=1}^{3} \sigma_{k,m} \leq 1$ for $k = 1, 2, 3$, and $\sum_{m=1}^{3} \sigma_{m,\ell} \leq 1$ for $\ell = 1, 2, 3$. Figure 1 shows two possible matchings. On the left-hand side, the matching allows a packet to be transmitted from input port 3 to output port 2, but since $Q_{3,2}$ is empty, no packet is actually transmitted.

In general, for an n-port switch, there are $N = n^2$ queues. The corresponding schedule set $\mathcal{S}$ is defined as

$$(12) \qquad \mathcal{S} = \left\{ \sigma \in \{0, 1\}^{n\times n} : \sum_{m=1}^{n} \sigma_{k,m} \leq 1, \sum_{m=1}^{n} \sigma_{m,\ell} \leq 1, 1 \leq k, \ell \leq n \right\}.$$

It can be checked that $\mathcal{S}$ is *monotone*. Furthermore, due to Birkhoff–von Neumann theorem, [2, 37], the convex hull of $\mathcal{S}$ is given by

$$(13) \qquad \langle \mathcal{S} \rangle = \left\{ \mathbf{x} \in [0, 1]^{n\times n} : \sum_{m=1}^{n} x_{k,m} \leq 1, \sum_{m=1}^{n} x_{m,\ell} \leq 1, 1 \leq k, \ell \leq n \right\}.$$

Thus, the rank of $\langle \mathcal{S} \rangle$ is less than or equal to $2n = 2\sqrt{N}$ for an n-port switch. Finally, given an arrival rate matrix[3] $\lambda \in [0, 1]^{n\times n}$, $\rho(\lambda)$ is given by

$$\rho(\lambda) = \max_{1 \leq k, \ell \leq n} \left\{ \sum_{m=1}^{n} \lambda_{k,m}, \sum_{m=1}^{n} \lambda_{m,\ell} \right\}.$$

3. Related works. The question of determining the optimal scaling of queue sizes in switched networks, or more generally, stochastic processing networks, has

---

[3]Not a vector, for notational convenience, as discussed earlier.

been an important intellectual pursuit for more than a decade. The complexity of the generic stochastic processing network makes this task extremely challenging. Therefore, in search of tractable analysis, most of the prior work has been on trying to understand optimal scaling and scheduling policies for *scaled* systems: primarily, with respect to fluid and heavy-traffic scaling, that is, $\rho \to 1$.

In heavy-traffic analysis, one studies the queue-size behavior under a diffusion (or heavy-traffic) scaling. This regime was first considered by Kingman [21]; since then, a substantial body of theory has developed, and modern treatments can be found in [5, 14, 39, 40]. Stolyar [33] has studied a class of myopic scheduling policies, known as the maximum weight policy, introduced by Tassiulas and Ephremides [35], for a generalized switch model in the diffusion scaling. In a general version of the maximum weight policy, a schedule with maximum weight is chosen at each time step, with the weight of a schedule being equal to the sum of the weights of the queues chosen by that schedule. The weight of a queue is a function of its size. In particular, for the choice of one parameter class of functions parameterized by $\alpha > 0$, $f(x) = x^\alpha$, the resulting class of policies are called the maximum weight policies with parameter $\alpha > 0$, and denoted as MW-α.

In [33], a complete characterization of the diffusion approximation for the queue-size process was obtained, under a condition known as "*complete resource pooling*," when the network is operating under the MW-α policy, for any $\alpha > 0$. Stolyar [33] showed the remarkable result that the limiting queue-size vector lives in a one-dimensional state space. Operationally, this means that all one needs to keep track of is the one-dimensional total amount of work in the system (called the *rescaled workload*), and at any point in time one can assume that the individual queues have all been balanced. Furthermore, it was established that a max-weight policy minimizes the rescaled workload induced by any policy under the heavy-traffic scaling (with complete resource pooling). Dai and Lin [6, 7] have established that a similar result holds (with complete resource pooling) in the more general setting of a stochastic processing network. In summary, under the complete resource pooling condition, the results in [6, 7, 33] imply that the performance of the maximum weight policy in an input-queued switch, or more generally in a stochastic processing network, is always optimal (in the diffusion limit, and when each queue size is appropriately weighted). These results suggest that the average total queue-size scales as $1/(1 - \rho)$ in the $\rho \to 1$ limit. However, such analyses do not capture the dependence on the network scheduling structure $\mathcal{S}$. Essentially, this is because the complete resource pooling condition reduces the system to a one-dimensional space (which may be highly dependent on a network's structure), and optimality results are then initially expressed with respect to this one-dimensional space.

Motivated to capture the dependence of the queue sizes on the network scheduling structure $\mathcal{S}$, a heavy-traffic analysis of switched networks with multiple bottlenecks (without resource pooling) was pursued by Shah and Wischik [32]. They

established the so-called multiplicative state space collapse, and identified a member, denoted by MW-0^+ (obtained by taking $\alpha \to 0$), of the class of maximum-weight policies as optimal with respect to a critical fluid model. In a more recent work, Shah and Wischik [31] established the optimality of MW-0^+ with respect to overloaded fluid models as well. However, this collection of works stops short of establishing optimality for diffusion scaled queue-size processes.

Finally, we take note of the work by Meyn [23], which establishes that a class of generalized maximum weight policies achieve logarithmic [in $1/(1 - \rho)$] regret with respect to an optimal policy under certain conditions.

In a related model—the bandwidth-sharing network model—Kang et al. [18] have established a diffusion approximation for the proportionally fair bandwidth allocation policy, assuming a technical "local traffic" condition, but without assuming complete resource pooling.[4] They show that the resulting diffusion approximation has a product-form stationary distribution. Shah, Tsitsiklis and Zhong [30] have recently established that this product-form stationary distribution is indeed the limit of the stationary distributions of the original stochastic model (an interchange-of-limits result). As a consequence, if one could utilize a scheduling policy in a switched network that corresponds to the proportionally fair policy, then the resulting diffusion approximation will have a product-form stationary distribution, as long as the effective network scheduling structure $\mathcal{S}$ (precisely $\langle \mathcal{S} \rangle$) satisfies the "local traffic condition." Now, proportional fairness is a continuous-time rate allocation policy that usually requires rate allocations that are a convex combination of multiple schedules. In a switched network, a policy must operate in discrete time and has to choose one schedule at any given time from a finite discrete set $\mathcal{S}$. For this reason, proportional fairness cannot be implemented directly. However, a natural randomized policy inspired by proportional fairness is likely to have the same diffusion approximation (since the fluid models would be identical, and the entire machinery of Kang et al. [18], building upon the work of Bramson [5] and Williams [40], relies on a fluid model). As a consequence, if $\mathcal{S}$ (more accurately, $\langle \mathcal{S} \rangle$) satisfies the "local traffic condition," then effectively the diffusion-scaled queue sizes would have a product-form stationary distribution, and would result in bounds similar to those implied by our results. In comparison, our results are nonasymptotic, in the sense that they hold for any admissible load, have a product-form structure, and do not require technical assumptions such as the "local traffic condition." Furthermore, such generality is needed because there are popular examples, such as the input-queued switch, that do *not* satisfy the "local traffic condition."

Another line of works—so-called large-deviations analysis—concerns exponentially decaying bounds on the tail probability of the steady-state distributions of

---

[4]Kang et al. [18] assume that critically loaded traffic is such that all the constraints are saturated simultaneously.

queue sizes. Venkataramanan and Lin [36] established that the maximum weight policy with weight parameter $\alpha > 0$, MW-α, optimizes the tail exponent of the $1 + \alpha$ norm of the queue-size vector. Stolyar [34] showed that a so-called "exponential rule" optimizes the tail exponent of the max norm of the queue-size vector. However, these works do not characterize the tail exponent explicitly. See [28] which has the best-known *explicit* bounds on the tail exponent.

In the context of input-queued switches, the example that has primarily motivated this work, the policy that we propose has the average total queue size bounded within factor 2 of the same quantity induced by *any* policy, in the heavy-traffic limit. Furthermore, this result does not require conditions like complete resource pooling. More generally, our policy provides nonasymptotic bounds on queue sizes for every arrival rate and switch size. The policy even admits exponential tail bounds with respect to the stationary distribution, and the exponent of these tail bounds is *optimal*. These results are significant improvements to the state-of-the-art bounds for best performing policies for input-queued switches. As noted in the Introduction, our bound on the average total queue size is $\sqrt{N}$ times better than the existing bound for the maximum-weight policy, and $\log N/(1 - \rho)$ times better than that for the batching policy in [25]. (Here N is the number of queues, and ρ the system load.) For further details of these results, see [29].

For a generic switched network, our policy induces average total queue size that scale linearly with the *rank* of $\langle \mathcal{S} \rangle$, under the diffusion scaling. This is in contrast to the best-known bounds, such as those for maximum weight policy, where the average queue-size scales as N, under the diffusion scaling. Therefore, whenever the *rank* of $\langle \mathcal{S} \rangle$ is smaller than N (the number of queues), our policy provides tighter bounds. Under our policy, queue sizes admit exponential tail bounds. The bound on the distribution of queue sizes under our policy leads to an explicit characterization of the tail exponent, which is optimal for a wide range of single-hop switched networks, including input-queued switches and the independent-set model of wireless networks, when the underlying interference graph is perfect.

4. Insensitivity in stochastic networks. This section recalls the background on insensitive stochastic networks that underlies the main results of this work. We shall focus on descriptions of the insensitive bandwidth allocation in so-called bandwidth-sharing networks operating in continuous time. Properties of these insensitive networks are provided in the Appendix.

We consider a bandwidth-sharing network operating in continuous time with capacity constraints. The particular bandwidth-sharing policy of interest is the store-and-forward allocation (SFA) mentioned earlier. We shall use the SFA as an idealized policy to design online scheduling policies for switched networks. We now describe the precise model, the SFA policy, and its performance properties.

Model. Let time be continuous and indexed by $t \in \mathbb{R}_+$. Consider a network with $J \geq 1$ resources indexed from $1, \ldots, J$. Let there be N routes, and suppose

that each *packet* on route i consumes an amount $R_{ji} \geq 0$ of resource j, for each $j \in \{1, 2, \ldots, J\}$. Let $\mathcal{K}$ be the set of all resource–route pairs (j, i) such that route i uses resource j, that is, $\mathcal{K} = \{(j, i) : R_{ji} > 0\}$. Without loss of generality, we assume that for each $i \in \{1, 2, \ldots, N\}$, $\sum_{j=1}^{J} R_{ji} > 0$. Let $\mathbf{R}$ be the $J \times N$ matrix with entries R_{ji}. Let $\mathbf{C} \in \mathbb{R}_+^J$ be a positive *capacity* vector with components C_j. For each route i, *packets* arrive as an independent Poisson process of rate λ_i. Packets arriving on route i require a unit amount of service, deterministically.

We denote the number of packets on route i at time t by $M_i(t)$, and define the queue-size vector at time t by $\mathbf{M}(t) = [M_i(t)]_{i=1}^N \in \mathbb{Z}_+^N$. Each packet gets service from the network at a rate determined according to a bandwidth-sharing policy. We also denote the total residual workload on route i at time t by $W_i(t)$, and let the vector of residual workload at time t be $\mathbf{W}(t) = [W_i(t)]_{i=1}^N$. Once a packet receives its total (unit) amount of service, it departs the network.

We consider online, myopic bandwidth allocations. That is, the bandwidth allocation at time t only depends on the queue-size vector $\mathbf{M}(t)$. When there are m_i packets on route i, that is, if the vector of packets is $\mathbf{m} = [m_i]_{i=1}^N$, let the total bandwidth allocated to route i be $\phi_i(\mathbf{m}) \in \mathbb{R}_+$. We consider a processor-sharing policy, so that each packet on route i is served at rate $\phi_i(\mathbf{m})/m_i$, if $m_i > 0$. If $m_i = 0$, let $\phi_i(\mathbf{m}) = 0$. If the *bandwidth vector* $\boldsymbol{\phi}(\mathbf{m}) = [\phi_i(\mathbf{m})]_{i=1}^N$ satisfies the capacity constraints

$$(14) \qquad \mathbf{R}\boldsymbol{\phi}(\mathbf{m}) \leq \mathbf{C} \qquad \text{componentwise}$$

for all $\mathbf{m} \in \mathbb{Z}_+^N$, then, in light of Definition 2.2, we say that $\boldsymbol{\phi}(\cdot)$ is an *admissible bandwidth allocation*. A Markovian description of the system is given by a process $\mathbf{Y}(t)$ which contains the queue-size vector $\mathbf{M}(t)$ along with the residual workloads of the set of packets on each route.

Now, on average, λ_i units of work arrive to route i per unit time. Therefore, in order for the Markov process $\mathbf{Y}(\cdot)$ to be positive (Harris) recurrent, it is necessary that

$$(15) \qquad \mathbf{R}\boldsymbol{\lambda} < \mathbf{C} \qquad \text{componentwise.}$$

All such $\boldsymbol{\lambda} = [\lambda_i]_{i=1}^N \in \mathbb{R}_+^N$ will be called *strictly admissible*, in the same spirit as strictly admissible vectors for a switched network. Similarly to the corresponding switched network, given $\boldsymbol{\lambda} \in \mathbb{R}_+^N$, we can define $\rho(\boldsymbol{\lambda})$, the load induced by $\boldsymbol{\lambda}$, using (10), as well as $\tilde{\rho}_j = (\sum_i R_{ji}\lambda_i)/C_j$. Then by Lemma 2.4, $\rho(\boldsymbol{\lambda}) = \max_j \tilde{\rho}_j$, where $\tilde{\rho}_j$ can be interpreted as the load induced by $\boldsymbol{\lambda}$ on resource j.

Store-and-forward allocation (*SFA*) *policy.* We describe the store-and-forward allocation policy that was first considered by Massoulié and later analyzed in the thesis of Proutière [26]. Bonald and Proutière [4] established that this policy induces product-form stationary distributions and is insensitive with respect to phase-type distributions. This policy is shown to be *insensitive* for general service time distributions, including the deterministic service considered here, by

Zachary [41]. The relation between this policy, the proportionally fair allocation, and multi-class queuing networks is discussed in depth by Walton [38] and Kelly, Massoulié and Walton [20]. The insensitivity property implies that the invariant measure of the process $\mathbf{M}(t)$ only depends on the parameters $\boldsymbol{\lambda} = [\lambda_i]_{i=1}^N \in \mathbb{R}_+^N$, and no other aspects of the stochastic description of the system.

We first give an informal motivation for SFA. SFA is closely related to quasi-reversible queuing networks. Consider a continuous-time multi-class queuing network (without scheduling constraints) consisting of processor sharing queues indexed by $j \in \{1, \ldots, J\}$ and job types indexed by the routes $i \in \{1, \ldots, N\}$. Each route i job has a service requirement R_{ji} at each queue j, and a fixed service capacity C_j is shared between jobs at the queue. Here each job will sequentially visit all the queues (so-called store-and-forward) and will visit each queue a fixed number of times. If we assume that jobs on each route arrive as a Poisson process, then the resulting queuing network will be stable for all strictly admissible arrival rates. Moreover, each stationary queue will be independent with a queue size that scales, with its load ρ, as $\rho/(1 - \rho)$. For further details, see Kelly [19]. So, assuming each queue has equal load, the total number of jobs within the network is of the order $J\rho/(1 - \rho)$. In other words, these networks have the stability and queue-size scaling that we require, but do not obey the necessary scheduling constraints (14). However, these networks do produce an admissible schedule on average. For this reason, we consider an SFA policy which, given the number of jobs on each route, allocates the average rate with which jobs are transferred through this multi-class network. Next, we describe this policy (using notation similar to those used in [20, 38]).

Given $\mathbf{m} \in \mathbb{Z}_+^N$, define

$$U(\mathbf{m}) = \left\{ \tilde{\mathbf{m}} = (\tilde{m}_{ji} : (j, i) \in \mathcal{K}) \in \mathbb{Z}_+^{|\mathcal{K}|} : \sum_{j \,:\, j \in i} \tilde{m}_{ji} = m_i, \text{ for all } 1 \le i \le N \right\}.$$

For $\mathbf{L} \in \mathbb{Z}_+^J$, we also define

$$V(\mathbf{L}) = \left\{ \tilde{\mathbf{m}} = (\tilde{m}_{ji} : (j, i) \in \mathcal{K}) \in \mathbb{Z}_+^{|\mathcal{K}|} : \sum_{i \,:\, i \ni j} \tilde{m}_{ji} = L_j, \text{ for all } 1 \le j \le J \right\}.$$

Here, by notation $j \in i$ (and $i \ni j$) we mean $R_{ji} > 0$. For each $\tilde{\mathbf{m}} \in U(\mathbf{m})$, we exploit notation somewhat and define $\tilde{m}_j = \sum_{i \,:\, j \in i} \tilde{m}_{ji}$, for all $j \le J$. Also define

$$\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} = \frac{\tilde{m}_j!}{\prod_{i \,:\, j \in i} (\tilde{m}_{ji}!)}.$$

For $\mathbf{m} \in \mathbb{Z}_+^N$, we define $\Phi(\mathbf{m})$ as

$$(16) \qquad \Phi(\mathbf{m}) = \sum_{\tilde{\mathbf{m}} \in U(\mathbf{m})} \prod_{j \in J} \left(\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} \prod_{i \,:\, j \in i} \left(\frac{R_{ji}}{C_j} \right)^{\tilde{m}_{ji}} \right).$$

We shall define $\Phi(\mathbf{m}) = 0$ if any of the components of $\mathbf{m}$ is negative. The store-and-forward allocation (SFA) assigns rates according to the function $\boldsymbol{\phi}: \mathbb{Z}_+^N \rightarrow \mathbb{R}_+^N$, so that for any $\mathbf{m} \in \mathbb{Z}_+^N$, $\boldsymbol{\phi}(\mathbf{m}) = (\phi_i(\mathbf{m}))_{i=1}^N$, with

$$(17) \qquad \phi_i(\mathbf{m}) = \frac{\Phi(\mathbf{m} - \mathbf{e}_i)}{\Phi(\mathbf{m})},$$

where, recalling that $\mathbf{m} - \mathbf{e}_i$ is the same as $\mathbf{m}$ at all but the ith component, its ith component equals $m_i - 1$. The bandwidth allocation $\boldsymbol{\phi}(\mathbf{m})$ is the stationary throughput of jobs on the routes of a multi-class queuing network (described above), conditional on there being $\mathbf{m}$ jobs on each route.

A priori it is not clear if the above described bandwidth allocation is even admissible, that is, satisfies (14). This can be argued as follows. The $\boldsymbol{\phi}(\mathbf{m})$ can be related to the stationary throughput of a multi-class network with a finite number of jobs, $\mathbf{m}$, on each route. Under this scenario (due to finite number of jobs), each queue must be stable. Therefore, the load on each queue, $\mathbf{R}\boldsymbol{\phi}(\mathbf{m})$, must be less than the overall system capacity $\mathbf{C}$. That is, the allocation is admissible. The precise argument along these lines is provided in, for example, [20], Corollary 2 and [38], Lemma 4.1.

The SFA induces a product-form invariant distribution for the number of packets waiting in the bandwidth-sharing network and is insensitive. We summarize this in the following result.

THEOREM 4.1. *Consider a bandwidth-sharing network with $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$. Under the SFA policy described above, the Markov process $\mathbf{Y}(t)$ is positive (Harris) recurrent, and $\mathbf{M}(t)$ has a unique stationary probability distribution $\boldsymbol{\pi}$ given by*

$$(18) \qquad \boldsymbol{\pi}(\mathbf{m}) = \frac{\Phi(\mathbf{m})}{\Phi} \prod_{i=1}^N \lambda_i^{m_i} \qquad \text{for all } \mathbf{m} \in \mathbb{Z}_+^N,$$

where

$$(19) \qquad \Phi = \prod_{j=1}^J \left(\frac{C_j}{C_j - \sum_{i : i \ni j} R_{ji} \lambda_i} \right)$$

is a normalizing factor. Furthermore, the steady-state residual workload of packets waiting in the network can be characterized as follows. First, the steady-state distribution of the residual workload of a packet is independent from $\boldsymbol{\pi}$. Second, in steady state, conditioned on the number of packets on each route of the network, the residual workload of each packet is uniformly distributed on $[0, 1]$, and is independent from the residual workloads of other packets.

Note that statements similar to Theorem 4.1 have appeared in other works, for example, [3], [38], Proposition 4.2, and [20]. Theorem 4.1 is a summary of these statements, and for completeness, it is proved in Appendix A.

The following property of the stationary distribution $\boldsymbol{\pi}$ described in Theorem 4.1 will be useful.

PROPOSITION 4.2. *Consider the setup of Theorem* 4.1, *and let* $\boldsymbol{\pi}$ *be described by* (18). *Define a measure* $\tilde{\boldsymbol{\pi}}$ *on* $\mathbb{Z}_+^{|\mathcal{K}|}$ *as follows: for* $\tilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}$,

$$(20) \qquad \tilde{\boldsymbol{\pi}}(\tilde{\mathbf{m}}) = \frac{1}{\Phi} \prod_{j=1}^{J} \left(\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} \prod_{i \,:\, j \in i} \left(\frac{R_{ji}\lambda_i}{C_j} \right)^{\tilde{m}_{ji}} \right).$$

Then, for any $L \in \mathbb{Z}_+$,

$$(21) \qquad \boldsymbol{\pi}\left(\left\{ \mathbf{m} : \sum_{i=1}^{N} m_i = L \right\} \right) = \tilde{\boldsymbol{\pi}}\left(\left\{ \tilde{\mathbf{m}} : \sum_{j=1}^{J} \tilde{m}_j = L \right\} \right).$$

We relate the distribution $\tilde{\boldsymbol{\pi}}$ to the stationary distribution of an insensitive multi-class queuing network with a product-form stationary distribution and geometrically distributed queue sizes.

PROPOSITION 4.3. *Consider the distribution* $\tilde{\boldsymbol{\pi}}$ *defined in* (20). *Then, for any* $\mathbf{L} = (L_1, \ldots, L_J) \in \mathbb{Z}_+^J$,

$$(22) \qquad \tilde{\boldsymbol{\pi}}(\tilde{m}_1 = L_1, \ldots, \tilde{m}_J = L_J) = \sum_{(\tilde{m}_{ji}) \in V(\mathbf{L})} \tilde{\boldsymbol{\pi}}\big((\tilde{m}_{ji})\big) = \prod_{j=1}^{J} \tilde{\rho}_j^{L_j}(1 - \tilde{\rho}_j),$$

where $\tilde{\rho}_j = (\sum_{i \,:\, i \ni j} R_{ji}\lambda_j)/C_j$.

Using Theorem 4.1 and Propositions 4.2 and 4.3, we can compute the expected value and the probability tail exponent of the steady-state total residual workload in the system. Recall that the total residual workload in the system at time t is given by $\sum_{i=1}^{N} W_i(t)$.

PROPOSITION 4.4. *Consider a bandwidth-sharing network with* $\mathbf{R}\lambda < \mathbf{C}$, *operating under the SFA policy. Denote the load induced by* λ *to be* $\rho = \rho(\lambda)(< 1)$, *and for each* j, *let* $\tilde{\rho}_j = (\sum_i R_{ji}\lambda_i)/C_j$. *Then* $\mathbf{W}(\cdot)$ *has a unique stationary probability distribution. With respect to this stationary distribution, the following properties hold*:

(i) *The expected total residual workload is given by*

$$(23) \qquad \mathbb{E}\left[\sum_{i=1}^{N} W_i \right] = \frac{1}{2} \sum_{j=1}^{J} \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j}.$$

(ii) *The distribution of the total residual workload has an exponential tail with exponent given by*

$$(24) \qquad \lim_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{i=1}^{N} W_i \geq L\right) = -\theta^*,$$

where θ^* is the unique positive *solution of the equation* $\rho(e^{\theta} - 1) = \theta$.

5. Main result: A policy and its performance. In this section, we describe an online scheduling policy and quantify its performance in terms of explicit, closed-form bounds on the stationary distribution of the induced queue sizes. Section 5.1 describes the policy for a generic switched network and provides the statement of the main result. Section 5.2 discusses its implications. Specifically, it discusses (a) the optimality of the policy for a large class of switched networks with respect to exponential tail bounds, and (b) the optimality of the policy for a class of switched networks, including input-queued switches, with respect to the average total queue size. Section 5.3 proves the main result stated in Section 5.1.

5.1. *A policy for switched networks.* The basic idea behind the policy, to be described in detail shortly, is as follows. Given a switched network, denoted by **SN**, with constraint set $\mathcal{S}$ and N queues, let $\langle \mathcal{S} \rangle$ have rank J and representation [cf. (6)]

$$\langle \mathcal{S} \rangle = \{\mathbf{x} \in [0, 1]^N : \mathbf{R}\mathbf{x} \leq \mathbf{C}\}, \qquad \mathbf{R} \in \mathbb{R}_+^{N \times J}, \mathbf{C} \in \mathbb{R}_+^J.$$

Now consider a virtual bandwidth-sharing network, denoted by **BN**, with N routes corresponding to each of these N queues. The resource–route relation is determined precisely by the matrix $\mathbf{R}$, and the J resources have capacities given by $\mathbf{C}$. Both networks, **SN** and **BN**, are fed identical arrivals. That is, whenever a packet arrives to queue i in **SN**, a packet is added to route i in **BN** at the same time. The main question is that of determining a scheduling policy for **SN**; this will be derived from **BN**. Specifically, **BN** will operate under the insensitive SFA policy described in Section 4. By Theorem 4.1 and Propositions 4.2 and 4.3, this will induce a desirable stationary distribution of queue sizes in **BN**. Therefore, if we could use the rate allocation of **BN**, that is, the SFA policy, directly in **SN**, it would give us a desired performance in terms of the stationary distribution of the induced queue sizes. Now the rate allocation in **BN** is such that the instantaneous rate is always inside $\langle \mathcal{S} \rangle$. However, it could change all the time and need not utilize points of $\mathcal{S}$ as rates. In contrast, in **SN** we require that the rate allocation can change only once per discrete time slot and it must always employ one of the generators of $\langle \mathcal{S} \rangle$, that is, a schedule from $\mathcal{S}$. The key to our policy is an effective way to emulate the rate allocation of **BN** under SFA (or for that matter, any admissible bandwidth allocation) by utilizing schedules from $\mathcal{S}$ in an online manner and with the discrete-time constraint. We will see shortly that this emulation policy relies on $\mathcal{S}$ being monotone; cf. Assumption 2.1.

To that end, we describe this emulation policy. Let us start by introducing some useful notation. Let $\mathbf{A}(\cdot) = (A_i(\cdot))$ be the vector of exogenous, independent Poisson processes according to which unit-sized packets arrive to both **BN** and **SN**, simultaneously. Recall that $A_i(\cdot)$ is a Poisson process with rate λ_i. Let $\mathbf{M}(t) = (M_i(t))$ denote the vector of numbers of packets waiting on the N routes in **BN** at time $t \geq 0$. In **BN**, the services are allocated according to the SFA policy described in Section 4. Let $\Lambda^{\mathrm{SFA}}(\cdot) = (\Lambda_i^{\mathrm{SFA}}(\cdot)) \in \mathbb{R}_+^N$ denote the cumulative amount of service allocated to the N routes in **BN** under the SFA policy: $\Lambda_i^{\mathrm{SFA}}(t)$ denotes the total amount of service allocated to all packets on route i during the interval $[0, t]$, for $t \geq 0$, with $\Lambda_i^{\mathrm{SFA}}(0) = 0$ for $1 \leq i \leq N$. By definition, all components of $\Lambda^{\mathrm{SFA}}(\cdot)$ are nondecreasing and Lipschitz continuous. Furthermore, $(\Lambda^{\mathrm{SFA}}(t + s) - \Lambda^{\mathrm{SFA}}(t))/s \in \langle \mathcal{S} \rangle$ for any $t \geq 0$ and $s > 0$. Recall that the (right-)derivative of $\Lambda^{\mathrm{SFA}}(\cdot)$ is determined by $\mathbf{M}(\cdot)$ through the function $\boldsymbol{\phi}(\cdot)$ as defined in (17).

Now we describe the scheduling policy for **SN** that will rely on $\Lambda^{\mathrm{SFA}}(\cdot)$. Let $\mathbf{B}(\tau) = (B_i(\tau))$ denote the cumulative amount of service allocated in **SN** by the scheduling policy up to time slot $\tau \geq 0$, with $\mathbf{B}(0) = \mathbf{0}$. The scheduling policy determines how $\mathbf{B}(\cdot)$ is updated. Let $\mathbf{Q}(\tau) = (Q_i(\tau))$ be the queue sizes measured at the end of time slot τ. Let service be provided according to the scheduling policy instantly at the beginning of a time slot. Thus, the scheduling policy decides the schedule $d\mathbf{B}(\tau) = \mathbf{B}(\tau + 1) - \mathbf{B}(\tau) \in \mathcal{S}$ at the very beginning of time slot $\tau + 1$. This decision is made as follows. Let $\mathbf{D}(\tau) = \Lambda^{\mathrm{SFA}}(\tau) - \mathbf{B}(\tau)$. We will see shortly that under our policy, $\mathbf{D}(\tau)$ is always nonnegative. This fact will be useful at various places, and in particular, for bounding the discrepancy between the continuous-time policy SFA and its discrete-time emulation. Let $\rho(\mathbf{D}(\tau))$ be the optimal objective value in the optimization problem PRIMAL($\mathbf{D}(\tau)$) defined in (7). In particular, there exists a nonnegative combination of schedules in $\mathcal{S}$ such that

$$(25) \qquad \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \boldsymbol{\sigma} \geq \mathbf{D}(\tau) \quad \text{and} \quad \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} = \rho(\mathbf{D}(\tau)).$$

We claim that in fact, we can find nonnegative numbers $\alpha_{\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} \in \mathcal{S}$, such that

$$(26) \qquad \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma} = \mathbf{D}(\tau) \quad \text{and} \quad \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} = \rho(\mathbf{D}(\tau)).$$

This is formalized in the following lemma.

LEMMA 5.1. *Let* $\mathbf{D} \in \mathbb{R}_+^N$ *be a nonnegative vector. Consider the static planning problem* PRIMAL($\mathbf{D}$) *defined in* (7). *Let the optimal objective value to* PRIMAL($\mathbf{D}$) *be* $\rho(\mathbf{D})$. *Then there exists* $\alpha_{\boldsymbol{\sigma}} \geq 0$, $\boldsymbol{\sigma} \in \mathcal{S}$, *such that* (26) *holds.*

The proof of the lemma relies on Assumption 2.1, and is provided in the Appendix.

There could be many possible nonnegative combinations of $\mathbf{D}(\tau)$ satisfying (26). If there exist nonnegative numbers $\alpha_{\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} \in \mathcal{S}$, satisfying (26) with $\alpha_{\boldsymbol{\sigma}'} \geq 1$ for some $\boldsymbol{\sigma}' \in \mathcal{S}$, then choose $\boldsymbol{\sigma}'$ as the schedule: set $d\mathbf{B}(\tau) = \boldsymbol{\sigma}'$. If no such decomposition exists for $\mathbf{D}(\tau)$, then set $d\mathbf{B}(\tau) = \tilde{\boldsymbol{\sigma}}$, where $\tilde{\boldsymbol{\sigma}}$ is a solution (ties broken arbitrarily) of

$$(27) \qquad \text{maximize} \qquad \sum_i \sigma_i \quad \text{over} \quad \boldsymbol{\sigma} \in \mathcal{S}, \boldsymbol{\sigma} \leq \mathbf{D}(\tau).$$

Here first observe that for all time τ, $d\mathbf{B}(\tau) \leq \mathbf{D}(\tau)$, so $\mathbf{D}(\tau) \geq \mathbf{0}$. Hence, $\mathbf{0}$ is a feasible solution for the above problem, as $\mathbf{0} \in \mathcal{S}$.

The above is a complete description of the scheduling policy. Observe that it is an online policy, as the virtual network $\mathbf{BN}$ can be simulated in an online manner, and, given this, the scheduling decision in $\mathbf{SN}$ relies only on the history of $\mathbf{BN}$ and $\mathbf{SN}$. The following result quantifies the performance of the policy.

THEOREM 5.2. *Given a strictly admissible arrival rate vector* $\boldsymbol{\lambda}$, *with* $\rho = \rho(\boldsymbol{\lambda}) < 1$, *under the policy described above, the switched network* $\mathbf{SN}$ *is positive recurrent and has a unique stationary distribution. Let* $\tilde{\rho}_j = (\sum_i R_{ji}\lambda_i)/C_j$, $j = 1, 2, \ldots, J$ *be the same as in Proposition* 4.4. *With respect to this stationary distribution, the following properties hold*:

(1) *The expected total queue size is bounded as*

$$(28) \qquad \mathbb{E}\left[\sum_{i=1}^N Q_i\right] \leq \frac{1}{2}\left(\sum_{j=1}^J \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j}\right) + K(N + 2),$$

where $K = \max_{\boldsymbol{\sigma} \in \mathcal{S}}(\sum_i \sigma_i)$.

(2) *The distribution of the total queue size has an exponential tail with exponent given by*

$$(29) \qquad \lim_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{i=1}^N Q_i \geq L\right) = -\theta^*,$$

where θ^* *is the unique positive solution of the equation* $\rho(e^\theta - 1) = \theta$.

5.2. *Optimality of the policy.* This section establishes the optimality of our policy for input-queued switches, both with respect to expected total queue-size scaling and tail exponent. General conditions under which our policy is optimal with respect to tail exponent are also provided.

Scaling of queue sizes. We start by formalizing what we mean by the optimality of expected queue sizes and of their tail exponents. We consider policies under which there is a well-defined limiting stationary distribution of the queue sizes for all $\boldsymbol{\lambda}$ such that $\rho(\boldsymbol{\lambda}) < 1$. Note that the class of policies is not empty; indeed,

the maximum weight policy and our policy are members of this class. With some abuse of notation, let π denote the stationary distribution of the queue-size vector under the policy of interest. We are interested in two quantities:

(1) *Expected total queue size*. Let $\overline{Q}$ be the expected total queue size under the stationary distribution π, defined by

$$\overline{Q} = \mathbb{E}_\pi \left[\sum_i Q_i \right].$$

Note that by ergodicity, the time average of the total queue size and the expected total queue size under π are the same quantity.

(2) *Tail exponent*. Let $\beta_L(Q), \beta_U(Q) \in [-\infty, 0]$ be the lower and upper limits of the tail exponent of the total queue size under π (possibly $-\infty$ or 0), respectively, defined by

$$(30) \qquad \beta_L(Q) = \lim_{\ell \to \infty} \inf \frac{1}{\ell} \log \mathbb{P}_\pi \left(\sum_i Q_i \geq \ell \right) \quad \text{and}$$

$$(31) \qquad \beta_U(Q) = \lim_{\ell \to \infty} \sup \frac{1}{\ell} \log \mathbb{P}_\pi \left(\sum_i Q_i \geq \ell \right).$$

If $\beta_L(Q) = \beta_U(Q)$, then we denote this common value by $\beta(Q)$.

We are interested in policies that can achieve minimal $\overline{Q}$ and $\beta(Q)$. For tractability, we focus on *scalings* of these quantities with respect to $\mathcal{S}$ (equivalently, N) and $\rho(\lambda)$, as $1/(1 - \rho(\lambda))$ and N increase. For different λ' and λ, it is possible that $\rho(\lambda) = \rho(\lambda')$, but the scaling of $\overline{Q}$, for example, could be wildly different. For this reason, we consider the worst possible dependence on $1/(1 - \rho)$ and N among all λ with $\rho(\lambda) = \rho$.

Note that we are considering scalings with respect to two quantities, ρ and N, and we are interested in two limiting regimes, $\rho \to 1$ and $N \to \infty$. The optimality of queue-size scaling stated here is with respect to the order of limits $\rho \to 1$ and then $N \to \infty$. As noted in [29], taking the limits in different orders could potentially result in different limiting behaviors of the object of interest, for example, $\overline{Q}$. For further discussion, see Section 6. It should be noted, however, that whenever the tail exponent is optimal, this optimality holds for *any* ρ and N.

Optimality of the tail exponent.   Here we establish sufficient conditions under which our policy is optimal with respect to tail exponent. First, we present a universal lower bound on the tail exponent, for a general single-hop switched network under any policy. We then provide a condition under which this lower bound matches the tail exponent under our policy. This condition is satisfied by both input-queued switches and the independent-set model of wireless networks.

Consider any policy under which there exists a well-defined limiting stationary distribution of the queue sizes for all λ such that $\rho(\lambda) < 1$. Let π_0 denote

the stationary distribution of queue sizes under this policy. The following lemma establishes a universal lower bound on the tail exponent.

LEMMA 5.3. *Consider a switched network as described in Theorem* 5.2, *with scheduling set $\mathcal{S}$ and admissible region $\{\mathbf{x} \in [0, 1]^N : \mathbf{Rx} \leq \mathbf{C}\}$. Let $\boldsymbol{\pi}_0$ and $\boldsymbol{\lambda}$ be as described. For each j, let $\tilde{\rho}_j = \sum_{i=1}^N R_{ji}\lambda_i/C_j$ be defined as in Theorem* 5.2. *Then under $\boldsymbol{\pi}_0$,*

(32)
$$\liminf_{L \to \infty} \frac{1}{L} \log \mathbb{P}_{\boldsymbol{\pi}_0}\left(\sum_i Q_i \geq L\right) \geq - \min_{j=1,2,\ldots,J} \theta_j^*,$$

where, for each $j \in \{1, 2, \ldots, J\}$, θ_j^* is the unique positive solution of the equation

$$\sum_{i=1}^N \lambda_i\left(e^{R_{ji}\theta} - 1\right) = \theta.$$

PROOF. Consider a fixed $j \in \{1, 2, \ldots, J\}$. Without loss of generality, we assume that $C_j = 1$, by properly normalizing the inequality $(\mathbf{Rx})_j \leq C_j$. In this case, $R_{ji} \leq 1$ for all i, since for each $i \in \{1, 2, \ldots, N\}$, $\mathbf{e}_i \in \mathcal{S} \subset \langle\mathcal{S}\rangle$, and satisfies the constraint $(\mathbf{Re}_i)_j = R_{ji} \leq C_j = 1$.

Now consider the following single-server queuing system. The arrival process is given by the sum $\sum_{i=1}^N R_{ji} A_i(\cdot)$, so that arrivals across time slots are independent, and that in each time slot, the amount of work that arrives is $\sum_{i=1}^N R_{ji}a_i$, where a_i is an independent Poisson random variable with mean λ_i, for each i. Note that the arriving amount in a single time slot does not have to be integral. Note also that $\sum_{i=1}^N R_{ji}\lambda_i = \tilde{\rho}_j < 1$, since $\rho(\boldsymbol{\lambda}) = \max_j \tilde{\rho}_j < 1$. In each time slot, a unit amount of service is allocated to the total workload in the system. Then, for this system, the workload process $W(\cdot)$ satisfies

$$W(\tau + 1) = \left[W(\tau) - 1\right]^+ + \sum_{i=1}^N R_{ji}a_i(\tau),$$

where $a_i(\tau)$ is the number of arrivals to queue i in the original system in time slot τ. We make two observations for this system. First, $W(\cdot)$ is stochastically dominated by $\sum_{i=1}^N R_{ji}Q_i(\cdot)$, where $Q_i(\cdot)$ is the size of queue i in the original system, under any online scheduling policy. This is because for all schedules $\boldsymbol{\sigma} \in \mathcal{S}$, $\boldsymbol{\sigma}$ satisfies $\mathbf{R}\boldsymbol{\sigma} \leq \mathbf{C}$, and hence $\sum_{i=1}^N R_{ji}\sigma_i \leq C_j = 1$ for every $\boldsymbol{\sigma} \in \mathcal{S}$. Second, since $R_{ji} \leq 1$ for all i, $\sum_{i=1}^N R_{ji}Q_i(\cdot)$ is stochastically dominated by $\sum_{i=1}^N Q_i(\cdot)$. Thus we have

$$\liminf_{L \to \infty} \frac{1}{L} \log \mathbb{P}_{\boldsymbol{\pi}_0}\left(\sum_i Q_i \geq L\right) \geq \liminf_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(W(\infty) \geq L\right).$$

We now show that

$$\liminf_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(W(\infty) \geq L\right) \geq -\theta_j^*,$$

where θ_j^* is the unique positive solution of the equation

$$\sum_{i=1}^{N} \lambda_i \big(e^{R_{ji}\theta} - 1\big) = \theta.$$

Consider the log-moment generating function (log-MGF) of $\sum_{i=1}^{N} R_{ji}a_i$, the arriving amount in one time slot. Since a_i is a Poisson random variable with mean λ_i for each i, its moment generating function is given by

$$f(\theta) = \exp\left(\sum_{i=1}^{N} \lambda_i \big(e^{R_{ji}\theta} - 1\big)\right).$$

Hence the log-MGF is

$$\log f(\theta) = \sum_{i=1}^{N} \lambda_i \big(e^{R_{ji}\theta} - 1\big).$$

By Theorem 1.4 of [13],

$$\lim_{L \to \infty} \frac{1}{L} \log \mathbb{P}\big(W(\infty) \geq L\big) = -\theta_j^*,$$

where $\theta_j^* = \sup\{\theta > 0 : \log f(\theta) < \theta\}$. Since $\log f(\theta) - \theta$ is strictly convex, θ_j^* satisfies

$$\sum_{i=1}^{N} \lambda_i \big(e^{R_{ji}\theta_j^*} - 1\big) = \theta_j^*.$$

$j \in \{1, 2, \ldots, J\}$ is arbitrary, so

$$\liminf_{L \to \infty} \frac{1}{L} \log \mathbb{P}_{\pi_0}\left(\sum_i Q_i \geq L\right) \geq - \min_{j=1,2,\ldots,J} \theta_j^*. \qquad \square$$

For general switched networks, the lower bound above need not match the tail exponent achieved under our policy [cf. (29)]. However, for a wide class of switched networks, these two quantities are equal. The following corollary of Lemma 5.3 is immediate.

COROLLARY 5.4. *Consider a switched network as described in Lemma* 5.3, *with scheduling set* $\mathcal{S}$ *and admissible region* $\{\mathbf{x} \in [0, 1]^N : \mathbf{Rx} \leq \mathbf{1}\}$. *If for all* j *and* i, $R_{ji} \in \{0, 1\}$, *then our policy achieves optimal tail exponent, for any strictly admissible arrival-rate vector* $\boldsymbol{\lambda}$.

PROOF. Let $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ be strictly admissible, that is, $\mathbf{R}\boldsymbol{\lambda} < \mathbf{1}$. Let $\tilde{\rho}_j = \sum_i R_{ji}\lambda_i$ for each j, and let $\rho = \rho(\boldsymbol{\lambda})$ be the system load induced by $\boldsymbol{\lambda}$. Consider the θ_j^* in

Lemma 5.3. When $R_{ji} \in \{0, 1\}$ for all j, and i, θ_j^* is the unique positive solution of the equation

$$\tilde{\rho}_j (e^\theta - 1) = \theta$$

for each j. Using the relation $\rho = \max_j \tilde{\rho}_j$, we see that $\min_j \theta_j^*$ is the unique positive solution of the equation

$$\rho(e^\theta - 1) = \theta.$$

Comparing this with equation (29) of Theorem 5.2, we see that our policy achieves the optimal tail exponent. $\square$

Consider an $n \times n$ input-queued switch, defined in Section 2.4, and with the admissible region described by (13). By Corollary 5.4, it is clear that the tail exponent in input-queued switches is optimal under our policy. Moreover, input-queued switches are not the only network model that satisfies the condition stated in Corollary 5.4. For example, consider the independent-set model of a wireless network. When the underlying interference graph is bipartite, it is easy to see that the admissible region is characterized by inequalities of the form $x_i + x_j \leq 1$ over all edges (i, j) of the graph, and $x_i \leq 1$ for isolated nodes i. More generally, when the underlying graph is perfect, inequality constraints characterizing the admissible region take the form $\sum_i x_i \leq 1$, where the summation is over all vertices of a clique. This latter fact follows from a proof of the weak perfect graph theorem, see, for example, Theorem 12.1.2 in [22]. Thus the incidence matrix $\mathbf{R}$ has all entries in $\{0, 1\}$, and the tail exponent under our policy is optimal for this model.

Optimality in input-queued switches. Here we argue the optimality of our policy for input-queued switches. As discussed above, the scaling of tail exponent is optimal under our policy for input-queued switches. We would argue the optimal scaling of the average total queue size under our policy for input-queued switches. To that end, as argued in Shah, Tsitsiklis and Zhong [29], when all input and output ports approach critical load, the average total queue size under any policy for input-queued switch must scale at least as fast as $\sqrt{N}/(1 - \rho)$, for any n-port switch with $N = n^2$ queues. For completeness, we include the proof for this lower bound here. As in Section 2.4, we use double indexing.

LEMMA 5.5. *Consider an n-port input-queued switch, with an arrival rate vector* $\boldsymbol{\lambda}$. *Suppose that the loads on all input and output ports are* ρ, *that is,* $\sum_{k=1}^n \lambda_{k,\ell} = \sum_m \lambda_{\ell,m} = \rho$, *for all* $\ell \in \{1, 2, \ldots, n\}$, *where* $\rho \in (0, 1)$. *Consider any policy under which the queue-size process has a well-defined limiting stationary distribution, and let this distribution be denoted by* $\boldsymbol{\pi}_0$. *Then under* $\boldsymbol{\pi}_0$, *we must have*

$$\mathbb{E}_{\boldsymbol{\pi}_0} \left[\sum_{k,\ell=1}^n Q_{k,\ell} \right] \geq \frac{n\rho}{2(1 - \rho)}.$$

D. SHAH, N. S. WALTON AND Y. ZHONG

PROOF.   We consider the sums of queue sizes at each output port, that is, the quantities $\sum_{k=1}^{n} Q_{k,\ell}$ for each $\ell \in \{1, 2, \ldots, n\}$. Since at most one packet can depart at each time slot, $\sum_{k=1}^{n} Q_{k,\ell}$ stochastically dominates the queue size in an $M/D/1$ system, with arrival rate ρ and deterministic service rate 1. Therefore, for each $\ell \in \{1, 2, \ldots, n\}$,

$$\mathbb{E}_{\pi_0}\left[\sum_{k=1}^{n} Q_{k,\ell}\right] \geq \frac{\rho}{2(1 - \rho)}.$$

Here, $\frac{\rho}{2(1-\rho)}$ is the expected queue size in steady state in an $M/D/1$ system. Summing over ℓ gives us the desired bound. $\square$

The optimality in terms of the average total queue size is a direct consequence of Theorem 5.2 and Lemma 5.5.

COROLLARY 5.6.   *Consider the same setup as in Lemma 5.5. Then in the heavy-traffic limit $\rho \to 1$, our policy is 2-optimal in terms of the average total queue size. More precisely, consider the expected total queue size in the diffusion scale in steady state, that is, $(1 - \rho)\overline{Q}$. Then*

$$\limsup_{\rho \to 1}(1 - \rho)\overline{Q} \leq n$$

under our policy, and

$$\liminf_{\rho \to 1}(1 - \rho)\overline{Q} \geq \frac{n}{2}$$

under any other policy.

PROOF.   Lemma 5.5 implies that

$$\liminf_{\rho \to 1}(1 - \rho)\overline{Q} \geq \frac{n}{2}$$

under any policy. For the upper bound, note that by Theorem 5.2, under our policy,

$$\overline{Q} \leq \frac{J}{2(1 - \rho)} + (N + 2)K.$$

For input-queued switches, $J \leq 2n$, as remarked in Section 5.2, $N = n^2$ and $K = n$. Therefore, we have that under our policy, the expected total queue size satisfies

(33) $$\overline{Q} \leq \frac{n}{1 - \rho} + (n^2 + 2)n.$$

Now consider the steady-state heavy-traffic scaling $(1 - \rho)\mathbf{Q}$. We have that

(34) $$(1 - \rho)\overline{Q} \leq n + (1 - \rho)(n^2 + 2)n.$$

The term $(1 - \rho)(n^2 + 2)n$ goes to zero as $\rho \to 1$, and hence under our policy,

$$\limsup_{\rho \to 1}(1 - \rho)\overline{Q} \leq n. \qquad \square$$

Our policy is not optimal in terms of the average total queue size, in general switched networks. In cases where $J \gg N$, the moment bounds for the maximum-weight policy gives tighter upper bounds. For further discussion, see Section 6.

5.3. *Proof of Theorem* 5.2. The proof is divided into three parts. The first part describes a sample-path-wise relation between $\mathbf{Q}(\cdot)$ and $\mathbf{W}(\cdot)$, the residual work-load vector in **BN**, which states that $\mathbf{Q}(\cdot)$ and $\mathbf{W}(\cdot)$ differ only by at most a constant at all times. Note that this domination is a distribution-free statement. The second part utilizes this fact to establish the positive recurrence of the **SN** Markov chain. The third part, as a consequence of the first two parts, and using Theorem 4.1, establishes the quantitative claims in Theorem 5.2.

Part 1. *Dominance.* We start by establishing that the queue sizes $\mathbf{Q}(\cdot)$ of **SN** are effectively dominated by the workloads $\mathbf{W}(\cdot)$ of **BN** at all times. We state this result formally in Proposition 5.9, which is a consequence of Lemmas 5.7 and 5.8 below.

LEMMA 5.7. *Consider the evolution of queue sizes in both **BN** and **SN** networks fed by identical arrival process. Initially, $\mathbf{Q}(0) = \mathbf{M}(0) = \mathbf{0}$. Let $\mathbf{W}(\tau) = (W_i(\tau))$ denote the amount of unfinished work in all N queues under the **BN** network at time τ. Then for any $\tau \geq 0$ and $1 \leq i \leq N$,*

$$(35) \qquad W_i(\tau) \leq Q_i(\tau) \leq W_i(\tau) + D_i(\tau),$$

where $\mathbf{D}(\tau) = \Lambda^{\mathrm{SFA}}(\tau) - \mathbf{B}(\tau)$ is as described in Section 5.1.

PROOF. Consider any $i \in \{1, 2, \ldots, N\}$ and $\tau \geq 0$. From (4), in **SN**,

$$(36) \qquad Q_i(\tau) = A_i(\tau) - B_i(\tau) + Z_i(\tau),$$

where $Z_i(\tau)$ is the cumulative amount of idling at the ith queue in **SN**. Similarly in **BN**,

$$(37) \qquad W_i(\tau) = A_i(\tau) - \Lambda_i^{\mathrm{SFA}}(\tau) + \widehat{Z}_i(\tau),$$

where $\widehat{Z}_i(\tau)$ is the cumulative amount of idling for the ith queue in **BN**. Since by construction, $\mathbf{D}(\tau) = \Lambda^{\mathrm{SFA}}(\tau) - \mathbf{B}(\tau)$, and $\mathbf{D}(\tau) \geq \mathbf{0}$, we have that

$$(38) \qquad B_i(\tau) \leq \Lambda_i^{\mathrm{SFA}}(\tau) \leq B_i(\tau) + D_i(\tau).$$

By definition, the instantaneous rate allocation to the ith queue satisfies $\frac{d}{dt^+}\Lambda_i^{\text{SFA}}(t) = 0$ if $W_i(t) = 0$ [equivalently, if $M_i(t) = 0$] for any $t \geq 0$. Therefore, $\widehat{Z}_i(\tau) = 0$, and $W_i(\tau) = A_i(\tau) - \Lambda_i^{\text{SFA}}(\tau)$. On the other hand, by Skorohod's map,

$$
\begin{aligned}
Z_i(\tau) &= \sup_{0 \leq s \leq \tau} \left[B_i(s) - A_i(s)\right]^+ \\
&\leq \sup_{0 \leq s \leq \tau} \left[\Lambda_i^{\text{SFA}}(s) - A_i(s)\right]^+ \\
&= \widehat{Z}_i(\tau) = 0,
\end{aligned}
$$

(39)

hence for all i and τ, $Z_i(\tau) = 0$, and $Q_i(\tau) = A_i(\tau) - S_i(\tau)$. It then follows that

$$
\begin{aligned}
Q_i(\tau) &= A_i(\tau) - B_i(\tau) \\
&\leq A_i(\tau) - \Lambda_i^{\text{SFA}}(\tau) + D_i(\tau) \\
&= W_i(\tau) + D_i(\tau)
\end{aligned}
$$

(40)

and

$$
\begin{aligned}
W_i(\tau) &= A_i(\tau) - \Lambda_i^{\text{SFA}}(\tau) \\
&\leq A_i(\tau) - B_i(\tau) = Q_i(\tau).
\end{aligned}
$$

(41)

Inequalities (40) and (41) together imply (35). □

LEMMA 5.8. *Let* $\mathbf{D}(\tau)$ *be the same as in Lemma* 5.7. *Then, for all* $\tau \geq 0$, $\rho(\mathbf{D}(\tau)) \leq N + 2$. *In particular*,

$$
\sum_i D_i(\tau) \leq K(N+2) \qquad \text{where } K = \max_{\sigma \in \mathcal{S}} \sum_i \sigma_i.
$$

(42)

PROOF. This result is established as follows. First, observe that $\mathbf{D}(0) = \mathbf{0}$ and therefore $\rho(\mathbf{D}(0)) = 0$. Next, we show that $\rho(\mathbf{D}(\tau + 1)) \leq \rho(\mathbf{D}(\tau)) + 1$. That is, $\rho(\mathbf{D}(\cdot))$ can at most increase by 1 in each time slot. And finally, we show that $\rho(\mathbf{D}(\cdot))$ cannot increase once it exceeds $N + 1$. That is, if $\rho(\mathbf{D}(\tau)) \geq N + 1$, then $\rho(\mathbf{D}(\tau + 1)) \leq \rho(\mathbf{D}(\tau))$. This will complete the proof.

We start by establishing that $\rho(\mathbf{D}(\cdot))$ increases by at most 1 in unit time. By definition,

$$
\begin{aligned}
\mathbf{D}(\tau + 1) &= \Lambda^{\text{SFA}}(\tau + 1) - \mathbf{B}(\tau + 1) \\
&= \Lambda^{\text{SFA}}(\tau) - \mathbf{B}(\tau) + \left(\Lambda^{\text{SFA}}(\tau + 1) - \Lambda^{\text{SFA}}(\tau) - d\mathbf{B}(\tau)\right) \\
&= \mathbf{D}(\tau) + d\Lambda^{\text{SFA}}(\tau) - d\mathbf{B}(\tau) \\
&= \left(\mathbf{D}(\tau) - d\mathbf{B}(\tau)\right) + d\Lambda^{\text{SFA}}(\tau),
\end{aligned}
$$

(43)

where $d\Lambda^{\text{SFA}}(\tau) = \Lambda^{\text{SFA}}(\tau + 1) - \Lambda^{\text{SFA}}(\tau)$. As remarked earlier, $d\mathbf{B}(\tau) \leq \mathbf{D}(\tau)$ componentwise. Therefore, by (11) it follows that

$$\rho(\mathbf{D}(\tau + 1)) \leq \rho(\mathbf{D}(\tau) - d\mathbf{B}(\tau)) + \rho(d\Lambda^{\text{SFA}}(\tau)).$$

Note that $\rho(d\Lambda^{\text{SFA}}(\tau)) \leq 1$ because the instantaneous service rate under SFA is always admissible. Since $\mathbf{D}(\tau) \geq \mathbf{D}(\tau) - d\mathbf{B}(\tau) \geq \mathbf{0}$, any feasible solution to PRIMAL($\mathbf{D}(\tau)$) is also feasible to PRIMAL($\mathbf{D}(\tau) - d\mathbf{B}(\tau)$), and hence

$$\rho(\mathbf{D}(\tau) - d\mathbf{B}(\tau)) \leq \rho(\mathbf{D}(\tau)).$$

It follows that

(44) $$\rho(\mathbf{D}(\tau + 1)) \leq \rho(\mathbf{D}(\tau)) + 1.$$

Next, we shall argue that if $\rho(\mathbf{D}(\tau)) \geq N + 1$, then $\rho(\mathbf{D}(\tau + 1)) \leq \rho(\mathbf{D}(\tau))$. To that end, suppose that $\rho(\mathbf{D}(\tau)) \geq N + 1$. Now $\frac{1}{\rho(\mathbf{D}(\tau))}\mathbf{D}(\tau) \in \langle \mathcal{S} \rangle$. Note that $\langle \mathcal{S} \rangle$ is a convex set in a N-dimensional space with extreme points contained in $\mathcal{S}$. Therefore, by Carathéodory's theorem, $\frac{1}{\rho(\mathbf{D}(\tau))}\mathbf{D}(\tau)$ can be written as a convex combination of at most $N + 1$ elements in $\mathcal{S}$. That is, there exists $\alpha_k \geq 0$ with $\sum_{k=1}^{N+1} \alpha_k = 1$, and $\boldsymbol{\sigma}^k \in \mathcal{S}$, $k \in \{1, 2, \ldots, N + 1\}$, such that

(45) $$\frac{1}{\rho(\mathbf{D}(\tau))}\mathbf{D}(\tau) = \sum_{k=1}^{N+1} \alpha_k \boldsymbol{\sigma}^k.$$

Therefore, there exists some $k^* \in \{1, 2, \ldots, N + 1\}$, such that $\alpha_{k^*} \geq 1/(N + 1)$. Since $\rho(\mathbf{D}(\tau)) \geq N + 1$, $\rho(\mathbf{D}(\tau))\alpha_{k^*} \geq 1$. That is, $\mathbf{D}(\tau)$ can be written as a nonnegative combination of elements from $\mathcal{S}$ with one of them, $\boldsymbol{\sigma}^{k^*}$, having an associated coefficient that satisfies $\rho(\mathbf{D}(\tau))\alpha_{k^*} \geq 1$, as required. In this case, we have

(46) $$\mathbf{D}(\tau) - \boldsymbol{\sigma}^{k^*} = \sum_{k=1, k \neq k^*}^{N+1} \rho(\mathbf{D}(\tau))\alpha_k \boldsymbol{\sigma}^k + (\rho(\mathbf{D}(\tau))\alpha_{k^*} - 1)\boldsymbol{\sigma}^{k^*}.$$

Therefore,

(47) $$\rho(\mathbf{D}(\tau) - \boldsymbol{\sigma}^{k^*}) \leq \rho(\mathbf{D}(\tau)) - 1.$$

Our scheduling policy chooses such a schedule, that is, $\boldsymbol{\sigma}^{k^*}$; that is, $d\mathbf{B}(\tau) = \boldsymbol{\sigma}^{k^*}$. Therefore,

(48) $$\mathbf{D}(\tau + 1) = \mathbf{D}(\tau) - \boldsymbol{\sigma}^{k^*} + d\Lambda^{\text{SFA}}(\tau).$$

By another application of (11) it follows that

(49) $$\begin{aligned} \rho(\mathbf{D}(\tau + 1)) &\leq \rho(\mathbf{D}(\tau) - \boldsymbol{\sigma}^{k^*}) + \rho(d\Lambda^{\text{SFA}}(\tau)) \\ &\leq \rho(\mathbf{D}(\tau)) - 1 + 1, \\ &= \rho(\mathbf{D}(\tau)), \end{aligned}$$

where again we have used the fact that $\rho(d\Lambda^{\text{SFA}}(\tau)) \leq 1$, due to the feasibility of SFA policy and (47). This establishes that $\rho(\mathbf{D}(\tau)) \leq N + 2$ for all $\tau \geq 0$. That is, for each $\tau \geq 0$, there exists $\alpha_{\boldsymbol{\sigma}} \geq 0$ for all $\boldsymbol{\sigma} \in \mathcal{S}$, $\sum_{\boldsymbol{\sigma}} \rho(\mathbf{D}(\tau))\alpha_{\boldsymbol{\sigma}} \leq N + 2$ and

$$(50) \qquad \mathbf{D}(\tau) \leq \sum_{\boldsymbol{\sigma}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}.$$

Therefore,

$$\sum_i D_i(\tau) = \mathbf{D}(\tau) \cdot \mathbf{1} \leq \sum_{\boldsymbol{\sigma}} \rho(\mathbf{D}(\tau))\alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma} \cdot \mathbf{1}$$

$$(51) \qquad \leq \left(\sum_{\boldsymbol{\sigma}} \rho(\mathbf{D}(\tau))\alpha_{\boldsymbol{\sigma}} \right) \left(\max_{\boldsymbol{\sigma} \in \mathcal{S}} \sum_i \sigma_i \right) \leq (N + 2)K,$$

where $K = \max_{\boldsymbol{\sigma} \in \mathcal{S}} \sum_i \sigma_i$. This completes the proof of Lemma 5.8. $\quad\square$

Lemmas 5.7 and 5.8 together imply the following proposition.

PROPOSITION 5.9. *Let* $\mathbf{Q}(\cdot)$, $\mathbf{W}(\cdot)$ *and* $\mathbf{M}(\cdot)$ *be as in Lemma* 5.7. *Then*

$$(52) \qquad \sum_{i=1}^N Q_i(\tau) \leq \sum_{i=1}^N W_i(\tau) + K(N + 2) \leq \sum_{i=1}^N M_i(\tau) + K(N + 2),$$

where $K = \max_{\boldsymbol{\sigma} \in \mathcal{S}} (\sum_{i=1}^N \sigma_i)$.

PROOF. We obtain the bounds of (52) by summing inequality (35) over $i \in \{1, 2, \ldots, N\}$, and using bound (42). $\quad\square$

Part 2. *Positive recurrence.* We start by defining the Markov chain describing the system evolution under the policy of interest. There are essentially two systems that evolve in a coupled manner under our policy: the virtual bandwidth-sharing network **BN** and the switched network **SN** of interest. These two networks are fed by the same arrival processes which are exogenous and Poisson (and hence Markov). The virtual system **BN** has a Markovian state consisting of the packets whose services are not completed, represented by the vector $\mathbf{M}(\cdot)$, and their residual services. The residual services of $M_i(\cdot)$ packets queued on route i can be represented by a nonnegative, finite measure $\mu_i(\cdot)$ on $[0, 1]$: unit mass is placed at each of the points $0 \leq s_1, \ldots, s_{M_i(t)} \leq 1$ if the unfinished work of $M_i(t)$ packets are given by $0 < s_1, \ldots, s_{M_i(t)} \leq 1$.

We now consider a Markovian description of the network **SN** in discrete time: let $\mathbf{X}(\tau)$ be the state of the system defined as

$$(53) \qquad \mathbf{X}(\tau) = \big(\mathbf{M}(\tau), \boldsymbol{\mu}(\tau), \mathbf{Q}(\tau), \mathbf{D}(\tau) \big),$$

where $(\mathbf{M}(\tau), \boldsymbol{\mu}(\tau))$ represents the state of **BN** at time τ, $\mathbf{Q}(\tau)$ is the vector of queue sizes in **SN** at time τ and $\mathbf{D}(\tau)$ is the "difference" vector maintained by

the scheduling policy for **SN**, as described in Section 5.1. Observe that the state $\mathbf{X}(\tau + 1)$ is a function of the previous state $\mathbf{X}(\tau)$ and the independent random arrival times occurring in the time interval $(\tau, \tau + 1)$ according to our Poisson process. This ensures conditional independence between $\mathbf{X}(\tau + 1)$ and $\mathbf{X}(\tau - 1)$ given $\mathbf{X}(\tau)$. So, by standard arguments, $\mathbf{X}(\cdot)$ is Markov and, indeed, strong Markov.

We now define the state space X of the Markov chain $\mathbf{X}(\cdot)$:

$$\mathsf{X} = \mathbb{Z}_+^N \times \mathcal{M}([0, 1])^N \times \mathbb{Z}_+^N \times \mathcal{D},$$

where $\mathcal{M}([0, 1])$ is the space of all nonnegative, finite measures on $[0, 1]$ and where $\mathcal{D} = (N + 2) \cdot \mathcal{C}$ is the admissible region $\mathcal{C}$ expanded by a multiplicative factor $N + 2$. The set $\mathcal{D}$ is exactly the set of vectors $\mathbf{d} \in \mathbf{R}_+^N$ for which $\rho(\mathbf{d}) \leq N + 2$; cf. (10). Thus, by Lemma 5.8, our process $\mathbf{D}(\cdot)$ can never leave the set $\mathcal{D}$.

We endow $\mathcal{M}([0, 1])$ with the weak topology, which is induced by the Prohorov's metric. This results in a complete and separable metric (Polish) space. The set $\mathcal{D}$ is a closed convex subset of $\mathbf{R}_+^N$. We endow $\mathbb{Z}_+$ and $\mathcal{D}$ with the obvious metrics (e.g., ℓ_1). The entire product space is endowed with the metric that is the maximum of metrics on component spaces. The resulting product space is Polish, on which a Borel σ-algebra, $\mathcal{B}_{\mathsf{X}}$, can be defined.

We remark that the Markov chain $\mathbf{X}(\cdot)$ need not be recurrent (nor neighborhood recurrent) for all states in X. However, we can start our Markov chain from any state $x \in \mathsf{X}$ and it will hit state $\mathbf{0}$ in finite expected time. We can then prove that our Markov chain is positive Harris recurrent. The resulting stationary measure defines the subset of X for which $\mathbf{X}(\cdot)$ is recurrent.

Given the Markovian description $\mathbf{X}(\tau)$ of **SN**, we establish its positive Harris recurrence in the following lemma.

LEMMA 5.10. *Consider a switched network SN with a strictly admissible arrival rate vector $\boldsymbol{\lambda}$, with $\rho(\boldsymbol{\lambda}) < 1$. Suppose that at time 0 the system is empty. Let $\mathbf{X}(\cdot)$ be as defined in equation (53). Then $\mathbf{X}(\cdot)$ is positive Harris recurrent and ergodic.*

The proof of the lemma is technical, and is deferred to Appendix C. The idea is that the evolution of **BN** is not affected by **SN**, and that **BN** is, on its own, positive recurrent. Hence, starting from any initial state, the Markov process $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot))$ that describes the evolution of **BN**, reaches the null state, that is, $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot)) = \mathbf{0}$ at some finite expected time. Once **BN** reaches the null state, it stays at this state for an arbitrarily large amount of time with positive probability. By our policy, $\mathbf{Q}(\cdot)$ and $\mathbf{D}(\cdot)$ can be driven to $\mathbf{0}$ within this time interval. This establishes that $\mathbf{X}(\cdot)$ reaches the null state in finite expected time and that $\mathbf{X}(\cdot)$ is positive recurrent.

Part 3. *Completing the proof.* The positive recurrence of the Markov chain $\mathbf{X}(\cdot)$ implies that it possesses a unique stationary distribution. Let $\overline{W} =$

$\mathbb{E}_{\pi}[\sum_{i=1}^{N} W_i]$, where, similarly to Proposition 4.4, W_i is the steady-state work-load on queue i in **BN**. By ergodicity, the time average of the total queue size equals the expected total queue size in steady state, that is, $\overline{Q}$, and similarly for $\overline{W}$. Therefore, by Proposition 5.9,

$$\overline{Q} \leq \overline{W} + K(N + 2).$$

By Proposition 4.4,

$$\overline{W} = \frac{1}{2}\left(\sum_{j=1}^{J} \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j}\right).$$

Thus

$$\overline{Q} \leq \overline{W} + K(N + 2) = \frac{1}{2}\left(\sum_{j=1}^{J} \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j}\right) + K(N + 2).$$

We now establish the tail exponent in (29). By Proposition 5.9,

$$\sum_{i=1}^{N} W_i(\tau) \leq \sum_{i=1}^{N} Q_i(\tau) \leq \sum_{i=1}^{N} W_i(\tau) + K(N + 2),$$

deterministically and for all times τ. Since $K(N + 2)$ is a constant, $\sum_{i=1}^{N} Q_i(\cdot)$ and $\sum_{i=1}^{N} W_i(\cdot)$ have the same tail exponent in steady state. By Proposition 4.4, the tail exponent $\beta(\mathbf{W})$ of $\sum_{i=1}^{N} W_i$ in steady state is given by $-\theta^*$, where θ^* is the unique positive solution of the equation $\rho(e^{\theta} - 1) = \theta$, so

$$\beta(\mathbf{Q}) = \beta(\mathbf{W}) = -\theta^*.$$

6. Discussion. We present a novel scheduling policy for a general single-hop switched network model. The policy, in effect, emulates the so-called Store-and-forward (SFA) continuous-time bandwidth-sharing policy. The insensitivity property of SFA along with the relation of its stationary distribution with that of a multi-class queuing network leads to the explicit characterization of the stationary distribution of queue sizes induced by our policy. This allows us to establish the optimality of our policy in terms of tail exponent for a large class of switched networks, including input-queued switches, and the independent-set model of wireless networks when the underlying interference graph is perfect, and that with respect to the average total queue size for a class of switched networks, including the input-queued switches. As a consequence, this settles a conjecture stated in [29]. On the technical end, a key contribution of the paper is creating a discrete-time scheduling policy from a continuous-time rate allocation policy, and this may be of independent interest in other domains of applications. We also remark that the idea of designing a discrete-time policy by emulating a continuous-time policy is not new; for example, similar emulation schemes have appeared in [9, 12]. Our

emulation scheme is novel in that it captures the switched network structure where queues may be served simultaneously. This simultaneity of service is absent from earlier models.

The switched network model considered here requires the arrival processes to be Poisson. However, this is not a major restriction, due to a *Poissonization* trick considered, for example, in [10] and [17]: all arriving packets are first passed through a "regularizer," which emits out packets according to a Poisson process with a rate that lies between the arrival rate and the network capacity. This leads to the arrivals being effectively Poisson, as seen by the system with a somewhat higher rate—by choosing the rate of "regularizer" so that the effective gap to the capacity, that is, $(1 - \rho)$, is decreased by factor 2.

The scheduling policy that we propose is not optimal for general switched networks. For example, in the independent-set model of ad-hoc wireless networks, there are as many constraints as the number of edges in the interference graph, which is often much larger than the number of nodes. Under our policy, the average total queue size would scale with the number of edges, whereas maximum-weight policy achieves a scaling with the number of nodes.

There are many possible directions for future research. One direction is the search for low-complexity and optimal scheduling policies. In the context of input-queued switches, our policy has a complexity that is exponential in N, the number of queues, because one has to compute the sum of exponentially many terms at every time instance. This begs the question of finding an optimal policy with polynomial complexity in N. One candidate is the MW-α policy, $\alpha > 0$, which has polynomial complexity, but its optimality appears difficult to analyze. Another possible candidate could be, as discussed in the Introduction, a randomized version of proportional fairness. The relationship between SFA and proportional fairness is explored in [38], where it was formally established that SFA converges to proportional fairness under the heavy-traffic limit, in an appropriate sense. The question remains whether (a version of) proportional fairness is optimal for input-queued switches.

Another interesting direction to pursue has to do the analysis of different limiting regimes. We are interested in two limits: $N \to \infty$, and $\rho \to 1$, where N is the number of queues, and ρ is the system load. Again, take the example of input-queued switches. In this paper, we have considered the heavy-traffic limit, that is, $\rho \to 1$, and show that our policy is optimal. However, if we take the limit $N \to \infty$, while keeping ρ fixed, then the average total queue-size scales as $N^{3/2}$, whereas maximum-weight policy produces a bound of N. A more interesting question is in the regime where $(1 - \rho)\sqrt{N}$ remain bounded, and where $N \to \infty$. In this regime, under our policy, under the maximum-weight policy, and under the batching policy in [25], the average total queue sizes all scale as $O(N^{3/2})$. In contrast, the scaling conjectured in [29] is $O(N)$. It is therefore of interest to see whether the 3/2 barrier can be broken. In [27], the authors device a policy that achieves N^{γ} scaling, for some $\gamma \in [1, 3/2)$.

APPENDIX A: PROPERTIES OF SFA

This section proves results stated in Section 4, specifically Theorem 4.1, Propositions 4.2, 4.3 and 4.4. First, we note that Propositions 4.2 and 4.3 are fairly easy consequences of Theorem 4.1, and their proofs are included for completeness. We then prove Proposition 4.4. Theorem 4.1 follows from the work of Zachary [41].

PROOF OF PROPOSITION 4.2. To verify (21), we can calculate both sides of the equation directly. Note that by definition, $\tilde{m}_j = \sum_{i \,:\, j \in i} \tilde{m}_{ji}$, so

$$(54) \qquad \tilde{\pi}\left(\left\{\tilde{\mathbf{m}} : \sum_{j=1}^{J} \tilde{m}_j = L\right\}\right) = \tilde{\pi}\left(\left\{\tilde{\mathbf{m}} : \sum_{(j,i) \in \mathcal{K}} \tilde{m}_{ji} = L\right\}\right).$$

On the other hand,

$$\pi\left(\left\{\mathbf{m} : \sum_{i=1}^{N} m_i = L\right\}\right)$$

$$(55) \qquad = \sum_{\mathbf{m} \in \mathbb{Z}_+^{|I|}} \mathbb{I}\left[\sum_{i=1}^{N} m_i = L\right] \frac{\Phi(\mathbf{m})}{\Phi} \prod_{i=1}^{N} \lambda_i^{m_i}$$

$$= \frac{1}{\Phi} \sum_{\mathbf{m} \in \mathbb{Z}_+^{|I|}} \mathbb{I}\left[\sum_{i=1}^{N} m_i = L\right]$$

$$(56) \qquad \times \sum_{\tilde{\mathbf{m}} \in U(\mathbf{m})} \prod_{i=1}^{N} \lambda_i^{m_i} \prod_{j=1}^{J}\left(\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} \prod_{i \,:\, j \in i}\left(\frac{R_{ji}}{C_j}\right)^{\tilde{m}_{ji}}\right)$$

$$= \frac{1}{\Phi} \sum_{\mathbf{m} \in \mathbb{Z}_+^{|I|}} \sum_{\tilde{\mathbf{m}} \in U(\mathbf{m})} \mathbb{I}\left[\sum_{i=1}^{N} m_i = L\right]$$

$$(57) \qquad \times \prod_{j=1}^{J}\left(\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} \prod_{i \,:\, j \in i}\left(\frac{R_{ji}\lambda_i}{C_j}\right)^{\tilde{m}_{ji}}\right)$$

$$= \frac{1}{\Phi} \sum_{\tilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}} \mathbb{I}\left[\sum_{(j,i) \in \mathcal{K}} \tilde{m}_{ji} = L\right]$$

$$(58) \qquad \times \prod_{j=1}^{J}\left(\binom{\tilde{m}_j}{\tilde{m}_{ji} : i \ni j} \prod_{i \,:\, j \in i}\left(\frac{R_{ji}\lambda_i}{C_j}\right)^{\tilde{m}_{ji}}\right)$$

$$(59) \qquad = \tilde{\pi}\left(\left\{\tilde{\mathbf{m}} : \sum_{(j,i) \in \mathcal{K}} \tilde{m}_{ji} = L\right\}\right).$$

Equality (55) follows from the definition of $\boldsymbol{\pi}$ given in (18), (56) follows from the definition of $\Phi(\mathbf{m})$ given in (16), (57) follows from the fact that for $\tilde{\mathbf{m}} \in U(\mathbf{m})$, $\sum_{j \,:\, j \in i} \tilde{m}_{ji} = m_i$ for all $i \in \mathcal{I}$, (58) follows from the fact that

$$\sum_{\tilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}, \mathbf{m} \in \mathbb{Z}_+^{|I|}} \mathbb{I}\left[\sum_{i=1}^N m_i = L, \sum_{j \,:\, j \in i} \tilde{m}_{ji} = m_i\right] = \mathbb{I}\left[\sum_{(j,i) \in \mathcal{K}} \tilde{m}_{ji} = L\right]$$

and (59) follows from the definition of $\tilde{\boldsymbol{\pi}}$ given in (20). Equations (54) and (59) together establish (21). $\quad\square$

PROOF OF PROPOSITION 4.3. We can verify (22) directly. Indeed,

$$\tilde{\boldsymbol{\pi}}\big(\{\tilde{m}_j = L_j : j = 1, 2, \ldots, J\}\big)$$

$$(60) \qquad = \frac{1}{\Phi} \sum_{\tilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}} \mathbb{I}\left[\sum_{i=1}^N \tilde{m}_{ji} = L_j\right] \prod_{j=1}^J \left(\binom{L_j}{\tilde{m}_{ji} : i \ni j} \prod_{i \,:\, j \in i} \left(\frac{R_{ji}\lambda_i}{C_j}\right)^{\tilde{m}_{ji}}\right)$$

$$(61) \qquad = \frac{1}{\Phi} \prod_{j=1}^J \left(\sum_{i \,:\, j \in i} \frac{R_{ji}\lambda_i}{C_j}\right)^{L_j}$$

$$(62) \qquad = \prod_{j=1}^J \left(\frac{C_j - \sum_{i \,:\, i \ni j} R_{ji}\lambda_i}{C_j}\right)\left(\sum_{i \,:\, j \in i} \frac{R_{ji}\lambda_i}{C_j}\right)^{L_j}$$

$$= \prod_{j=1}^J (1 - \tilde{\rho}_j)\tilde{\rho}_j^{L_j}.$$

Equality (60) follows from the definition of $\tilde{\boldsymbol{\pi}}$ in (20). Equality (61) collects all terms in the Newton expansion of the term $(\sum_{i \,:\, i \ni j} \frac{R_{ji}\lambda_i}{C_j})^{L_j}$. Equality (62) follows from the definition of Φ. $\quad\square$

PROOF OF PROPOSITION 4.4. Consider $\sum_{i=1}^N M_i$, the total number of packets waiting in **BN**, in steady state. By Propositions 4.2 and 4.3, $\sum_{i=1}^N M_i$ has the same distribution as the sum of J geometric random variables, with parameters $1 - \tilde{\rho}_1, \ldots, 1 - \tilde{\rho}_J$. Hence,

$$\mathbb{E}\left[\sum_{i=1}^N M_i\right] = \sum_{j=1}^J \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j}.$$

By Theorem 4.1, the individual residual workload in steady state is independent from the number of packets in the network, and is uniformly distributed on $[0, 1]$. Thus

$$\mathbb{E}\left[\sum_{i=1}^N W_i\right] = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^N M_i\right] = \frac{1}{2}\sum_{j=1}^J \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j}.$$

This establishes equation (23).

To establish equation (24), consider the following interpretation of $\sum_{i=1}^{N} W_i$, the total residual workload in steady state. By Theorem 4.1, $\sum_{i=1}^{N} W_i$ has the same distribution as $\sum_{\ell=1}^{M} U_\ell$, where $M = \sum_{i=1}^{N} M_i$, and U_ℓ are i.i.d. uniform random variables on $[0, 1]$, all independent from M. We first establish that

$$(63) \qquad \limsup_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{\ell=1}^{M} U_\ell \geq L \right) \leq -\theta^*,$$

where θ^* is the unique *positive* solution of the equation $\rho(e^\theta - 1) = \theta$. By Markov's inequality, for any $\theta > 0$, we have

$$\mathbb{P}\left(\sum_{\ell=1}^{M} U_\ell \geq L \right) \leq \exp(-\theta L) \mathbb{E}\left[\exp\left(\theta \sum_{\ell=1}^{M} U_\ell \right) \right]$$

$$= \exp(-\theta L) \mathbb{E}\left[\mathbb{E}\left[\exp\left(\theta \sum_{\ell=1}^{M} U_\ell \right) \Big| M \right] \right]$$

$$= \exp(-\theta L) \mathbb{E}\left[\left(\frac{e^\theta - 1}{\theta} \right)^M \right].$$

For notational convenience, let $x = \frac{e^\theta - 1}{\theta}$. We now consider the term $\mathbb{E}[x^M]$. Let $\widetilde{M}_j$ be independent geometric random variables with parameter $1 - \tilde{\rho}_j$, $j = 1, 2, \ldots, J$. Then M is distributed as $\sum_{j=1}^{J} \widetilde{M}_j$. Thus

$$\mathbb{E}[x^M] = \mathbb{E}[x^{\sum_{j=1}^{J} \widetilde{M}_j}] = \prod_{j=1}^{J} \mathbb{E}[x^{\widetilde{M}_j}] = \prod_{j=1}^{J} \frac{1 - \tilde{\rho}_j}{1 - \tilde{\rho}_j x}$$

for any $x > 0$ with $\rho x < 1$ (note that $\tilde{\rho}_j x < 1$ for all j if and only if $\rho x = \max_j \tilde{\rho}_j x < 1$, by Lemma 2.4). Therefore, for all $\theta > 0$ such that $\rho x = \rho(e^\theta - 1)/\theta < 1$, we have

$$\limsup_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{\ell=1}^{M} U_\ell \geq L \right) \leq \limsup_{L \to \infty} \frac{1}{L} \log \left\{ \exp(-\theta L) \prod_{j=1}^{J} \frac{1 - \tilde{\rho}_j}{1 - \tilde{\rho}_j x} \right\} = -\theta.$$

Taking the infimum over θ satisfying $\rho(e^\theta - 1)/\theta < 1$, we have established (63), that is,

$$\limsup_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{\ell=1}^{M} U_\ell \geq L \right) \leq -\theta^*.$$

We now prove the converse inequality. Without loss of generality, suppose that $\rho = \tilde{\rho}_1$, and $\widetilde{M}_1$ is a geometric random variable with parameter $1 - \rho$. Then we can

couple $\sum_{\ell=1}^{M} U_\ell$ and $\sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ on the same probability space so that $\sum_{\ell=1}^{M} U_\ell \geq \sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ with probability 1. Thus, it suffices to show that

$$\liminf_{L \to \infty} \frac{1}{L} \log \mathbb{P}\left(\sum_{\ell=1}^{\widetilde{M}_1} U_\ell \geq L \right) \geq -\theta^*.$$

Instead of calculating the quantity directly, consider a $M/D/1$ queue with load ρ, under the processor-sharing (PS) policy. Note that for this queuing system, SFA coincides with the PS policy. By Theorem 4.1, $\sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ is the steady-state distribution of the total residual workload in the system. On the other hand, consider the same queuing system under a FIFO policy. Since the workload is the same under any work-conserving policy, $\sum_{\ell=1}^{\widetilde{M}_1} U_\ell$ is also the steady-state distribution of the total workload in this system, which we denote by W_{FIFO}. By Theorem 1.4 of [13], we can characterize $\frac{1}{L} \log \mathbb{P}(W_{\text{FIFO}} \geq L)$ as follows. Let $f(\theta) = \log \mathbb{E}[e^{\theta X}]$, where X is a Poisson random variable with parameter ρ. Then we have

$$\lim_{L \to \infty} \frac{1}{L} \log \mathbb{P}(W_{\text{FIFO}} \geq L) = -\theta^*,$$

where $\theta^* = \sup\{\theta > 0 : f(\theta) < \theta\}$. It is a simple calculation to see that $f(\theta) = \rho(e^\theta - 1)$, so $\theta^* > 0$ satisfies $f(\theta^*) = \theta^*$. With this lower bound and the upper bound (63), we have established (24). $\square$

We now provide justifications for Theorem 4.1. Consider a bandwidth-sharing network model as described in Section 4. Instead of having packets requiring a unit amount of service, suppose each route i packet has a service requirement that is independent identically distributed with distribution μ_i and mean 1. We note that such bandwidth-sharing networks are a special case of the processor-sharing (PS) queuing network model, as considered by Zachary [41]. In particular, a bandwidth-sharing network is a processor-sharing network, where network jobs depart the network after completing service. General, insensitivity results for the bandwidth-sharing networks follow as a consequence of the work of Zachary [41].

Following Zachary [41], for $i \in \{1, 2, \ldots, N\}$, we define the probability distribution $\bar{\mu}_i$ to be the *stationary residual life distribution* of the renewal process with inter-event distribution μ_i. That is, if μ_i has cumulative distribution function F, then $\bar{\mu}_i$ has distribution function G given by

$$G(x) = 1 - \int_x^\infty (1 - F(y)) \, dy, \qquad x \geq 0.$$

Note that if the service requests are deterministically 1, that is, μ_i is the distribution of the deterministic constant 1, then $\bar{\mu}_i$ is a uniform distribution on $[0, 1]$, for all $i \in \{1, 2, \ldots, N\}$.

Insensitive rate allocation. Consider a bandwidth-sharing network described above, with rate allocation $\boldsymbol{\phi}(\cdot)$. A Markovian description of the system is given by a process $\mathbf{Y}(t)$ which contains the queue-size vector $\mathbf{M}(t)$ along with the residual workloads of the set of packets on each route. If the Markov process $\mathbf{Y}(t)$ admits an invariant measure, then it induces an invariant measure $\boldsymbol{\pi}$ on the process $\mathbf{M}(t)$. Such $\boldsymbol{\pi}$, when it exists, is called *insensitive* if it depends on the statistics of the arrivals and service requests only through the parameters $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^{N}$; in particular, it does not depend on the detailed service distributions of incoming packets. A rate allocation $\boldsymbol{\phi}(\cdot) = (\phi_i(\cdot))_{i=1}^{N}$ is called *insensitive* if it induces an insensitive invariant measure $\boldsymbol{\pi}$ on $\mathbf{M}(t)$.

It turns out that if the rate allocation $\boldsymbol{\phi}$ satisfies a *balance property*, then it is insensitive.

DEFINITION A.1 (Definition 1, [4]). Consider the bandwidth-sharing network just described. The rate allocation $\boldsymbol{\phi}(\cdot)$ is *balanced* if there exists a function $\Phi : \mathbb{Z}^N \to \mathbb{R}_+$ with $\Phi(\mathbf{0}) = 1$, and $\Phi(\mathbf{m}) = 0$ for all $\mathbf{m} \notin \mathbb{Z}_+^N$, such that

$$(64) \qquad \phi_i(\mathbf{m}) = \frac{\Phi(\mathbf{m} - \mathbf{e}_i)}{\Phi(\mathbf{m})} \qquad \text{for all } \mathbf{m} \in \mathbb{Z}_+^N, i \in \{1, 2, \ldots, N\}.$$

Bonald and Proutiére [3] proved that a balanced rate allocation is insensitive with respect to all phase-type service distributions. Zachary [41] showed that a balanced rate allocation is indeed insensitive with respect to all general service distributions. He also gave the characterization of the distribution of the residual workloads in steady state.

THEOREM A.2 (Theorem 2, [41]). *Consider the bandwidth-sharing network described earlier. A measure $\boldsymbol{\pi}$ on $\mathbb{Z}_+^N$ is stationary for $\mathbf{M}(t)$ and is insensitive to all service distributions with mean 1, if and only if it is related to the rate allocation $\boldsymbol{\phi}$ as follows*:

$$(65) \qquad \boldsymbol{\pi}(\mathbf{m})\phi_i(\mathbf{m}) = \boldsymbol{\pi}(\mathbf{m} - \mathbf{e}_i)\lambda_i \qquad \textit{for all } \mathbf{m} \in \mathbb{Z}_+^N, i \in \{1, 2, \ldots, N\},$$

where we set $\boldsymbol{\pi}(\mathbf{m} - \mathbf{e}_i)$ to be 0, if $m_i = 0$. Consequently, $\boldsymbol{\pi}$ is given by expression

$$(66) \qquad \boldsymbol{\pi}(\mathbf{m}) = \Phi(\mathbf{m}) \prod_{i=1}^{N} \lambda_i^{m_i}.$$

Furthermore, if $\boldsymbol{\pi}$ can be normalized to a probability distribution, then $\mathbf{Y}(t)$ is positive recurrent, and in steady state, the residual workload of each route-i packet in the network is distributed as $\bar{\mu}_i$, independent from $\boldsymbol{\pi}$, and is conditionally independent from the residual workloads of other packets, when we condition on the number of packets on each route of the network.

Note that conditions (64) and (65) are equivalent. Suppose that $\boldsymbol{\phi}(\cdot)$ satisfies (64), then an invariant measure $\boldsymbol{\pi}$ is given by (66). Substituting equation (66) into equation (64) gives equation (65). Conversely, if equation (65) is satisfied, then we can just set $\Phi(\mathbf{m}) = \boldsymbol{\pi}(\mathbf{m}) / \prod_{i=1}^{N} \lambda_i^{m_i}$, and equations (64) and (66) are satisfied.

PROOF OF THEOREM 4.1. Theorem 4.1 is now a fairly easy consequence of Theorem A.2. Consider a bandwidth-sharing network described in Section 4. The additional structures are the additional capacity constraints (14), and that arriving packets only require an unit amount of service, deterministically. The capacity constraints (14) impose the necessary condition for stability, given by (15). Recall that all arrival rate vectors $\boldsymbol{\lambda}$ that satisfy $\mathbf{R}\boldsymbol{\lambda} < \mathbf{C}$ are called *strictly admissible*.

Consider the bandwith vector $\boldsymbol{\phi}$ as defined by (16) and (17). As remarked earlier, $\boldsymbol{\phi}$ is admissible, that is, it satisfies the capacity constraints (14). It is balanced by definition, and hence is insensitive by Theorem A.2. Thus, it induces an stationary measure $\boldsymbol{\pi}$ on the queue-size vector $\mathbf{M}(t)$, given by (66). For a strictly admissible arrival rate vector $\boldsymbol{\lambda}$, the measure is finite, with the normalizing constant Φ given by (19). Hence, we can normalize $\boldsymbol{\pi}$ to obtain the unique stationary probability distribution for $\mathbf{M}(t)$.

Finally, using Theorem A.2 and the fact that all service requests are deterministically 1, we see that the stationary residual workloads are all uniformly distributed on $[0, 1]$ and independent. $\square$

APPENDIX B: PROOF OF LEMMA 5.1

Let $\mathbf{D} \geq \mathbf{0}$, and let $(\tilde{\alpha}_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma} \in \mathcal{S}}$ be an optimal solution to the program PRIMAL($\mathbf{D}$). Then $\tilde{\alpha}_{\boldsymbol{\sigma}} \geq 0$ for all $\boldsymbol{\sigma} \in \mathcal{S}$, $\mathbf{D} \leq \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \boldsymbol{\sigma}$, and $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} = \rho(\mathbf{D})$. We will construct $(\alpha_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma} \in \mathcal{S}}$ from $(\tilde{\alpha}_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma} \in \mathcal{S}}$ such that $\alpha_{\boldsymbol{\sigma}} \geq 0$ for all $\boldsymbol{\sigma} \in \mathcal{S}$, $\mathbf{D} = \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}$ and $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \alpha_{\boldsymbol{\sigma}} = \rho(\mathbf{D})$.

If $\mathbf{D} = \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \boldsymbol{\sigma}$, then there is nothing to prove. Thus, suppose that there exists i such that $D_i < \sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \sigma_i$. We now modify $(\tilde{\alpha}_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma} \in \mathcal{S}}$ to reduce the "gap" between $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \sigma_i$ and D_i.

Indeed, since $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \sigma_i > D_i \geq 0$, there is some $\tilde{\boldsymbol{\sigma}} \in \mathcal{S}$ such that $\tilde{\sigma}_i = 1$, and $\tilde{\alpha}_{\tilde{\boldsymbol{\sigma}}} > 0$. We now modify $(\tilde{\alpha}_{\boldsymbol{\sigma}})$ by reducing $\tilde{\alpha}_{\tilde{\boldsymbol{\sigma}}}$ by a positive amount

$$\varepsilon = \min(\tilde{\alpha}_{\tilde{\boldsymbol{\sigma}}}, \theta_i - D_i),$$

increasing $\tilde{\alpha}_{\tilde{\boldsymbol{\sigma}} - \mathbf{e}_i}$ by $\varepsilon > 0$, and keeping all other $\tilde{\alpha}_{\boldsymbol{\sigma}}$ the same ($\tilde{\boldsymbol{\sigma}} - \mathbf{e}_i \in \mathcal{S}$ by Assumption 2.1). Then it is easy to check that $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \sigma_i - D_i$ is reduced by ε, $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} \sigma_\ell - D_\ell$ remains the same for all $\ell \neq i$, $(\tilde{\alpha}_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma} \in \mathcal{S}}$ remain nonnegative and we still have $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} = \rho(\mathbf{D})$.

By repeating this procedure finitely many times, it follows that we can modify $(\tilde{\alpha}_{\boldsymbol{\sigma}})_{\boldsymbol{\sigma} \in \mathcal{S}}$ to make it satisfy (26). This completes the proof of Lemma 5.1.

APPENDIX C: PROOF OF LEMMA 5.10

First we note that by Theorem 4.1, **BN** is positive recurrent under the SFA policy, if $\rho(\boldsymbol{\lambda}) < 1$. Starting from any initial state, it also has a strictly positive probability of reaching the null-state $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot)) = \mathbf{0}$ at some finite time. Since the evolution of the virtual system **BN** does not depend on that of **SN**, it is, on its own, positive recurrent. Next we argue the positive recurrence of the entire network state building upon this property of **BN**.

Sufficient conditions to establish positive Harris recurrence and ergodicity of a discrete-time Markov chain $\mathbf{X}(\tau)$ with state space X are given by the following (see, [1], pages 198–202, and [11], Section 4.2, for details):

(C1) There exists a bounded set $A \in \mathcal{B}_{\mathsf{X}}$ such that

$$(67) \qquad \mathbb{P}_{\mathbf{x}}(T_A < \infty) = 1 \qquad \text{for any } \mathbf{x} \in \mathsf{X}$$

$$(68) \qquad \sup_{\mathbf{x} \in A} \mathbb{E}_{\mathbf{x}}[T_A] < \infty.$$

In the above, the stopping time $T_A = \inf\{\tau \geq 1 : \mathbf{X}(\tau) \in A\}$; notation $\mathbb{P}_{\mathbf{x}}(\cdot) \equiv \mathbb{P}(\cdot | \mathbf{X}(0) = \mathbf{x})$ and $\mathbb{E}_{\mathbf{x}}[\cdot] \equiv \mathbb{E}[\cdot | \mathbf{X}(0) = \mathbf{x}]$.

(C2) Given A satisfying (67)–(68), there exists $\mathbf{x}^* \in \mathsf{X}$, finite $\ell \geq 1$ and $\delta > 0$ such that

$$(69) \qquad \mathbb{P}_{\mathbf{x}}\big(\mathbf{X}(\ell) = \mathbf{x}^*\big) \geq \delta \qquad \text{for any } \mathbf{x} \in A$$

$$(70) \qquad \mathbb{P}_{\mathbf{x}^*}\big(\mathbf{X}(1) = \mathbf{x}^*\big) > 0.$$

Next, we verify conditions (C1) and (C2). For the set of points where **BN** is empty, say A, condition (C1) follows immediately from the following facts: (a) **BN** is positive recurrent and hence $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot))$ returns to $\mathbf{0}$ state in finite expected time starting from any finite state; (b) $\mathbf{D}(\cdot)$ is always bounded due to Lemma 5.8; and (c) $\mathbf{Q}(\cdot)$ returns to the bounded set $\sum_i Q_i(\cdot) \leq K(N+2)$ whenever $\mathbf{M}(\cdot) = \mathbf{0}$ due to Lemma 5.7. Condition (C2) can be verified for the null-state $\mathbf{x}^* = \mathbf{0}$ as follows: (a) $(\mathbf{M}(\cdot), \boldsymbol{\mu}(\cdot))$ returns to the null state with positive probability; (b) given this, it remains there for further $K(N+2)+1$ time with strictly positive probability due to Poisson arrival process; (c) in this additional time $K(N+2)+1$, the $\mathbf{Q}(\cdot)$ and $\mathbf{D}(\cdot)$ are driven to $\mathbf{0}$. To see (c), observe that when $\mathbf{M}(\cdot) = \mathbf{0}$, $\mathbf{D}(\cdot) \in \mathbb{Z}_+^N$. By construction of our policy and Assumption 2.1 on structure of $\mathcal{S}$, it follows that if $\mathbf{M}(\cdot)$ continues to remain $\mathbf{0}$, the $\sum_i D_i(\cdot)$ is reduced by at least unit amount till $\mathbf{D}(\cdot) = \mathbf{0}$; at which moment $\mathbf{Q}(\cdot)$ reaches $\mathbf{0}$ as well. Since $\sum_i D_i(\cdot) \leq K(N+2)$ by Lemma 5.8, it follows that $\mathbf{M}(\cdot)$ need to remain $\mathbf{0}$ for this to happen only for $K(N+2)+1$ amount of time. This completes the verification of the conditions (C1) and (C2). Consequently, we establish that the network Markov chain, represented by $\mathbf{X}(\cdot)$, is positive recurrent and ergodic.

REFERENCES

[1] ASMUSSEN, S. (2003). *Applied Probability and Queues*, 2nd ed. Springer, New York. MR1978607

[2] BIRKHOFF, G. (1946). Three observations on linear algebra. *Univ. Nac. Tucumán. Revista A*. **5** 147–151. MR0020547

[3] BONALD, T. and PROUTIÈRE, A. (2002). Insensitivity in processor-sharing networks. *Perform. Eval*. **49** 193–209.

[4] BONALD, T. and PROUTIÈRE, A. (2003). Insensitive bandwidth sharing in data networks. *Queueing Syst*. **44** 69–100. MR1989867

[5] BRAMSON, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst*. **30** 89–148. MR1663763

[6] DAI, J. G. and LIN, W. (2005). Maximum pressure policies in stochastic processing networks. *Oper. Res*. **53** 197–218. MR2131925

[7] DAI, J. G. and LIN, W. (2008). Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab*. **18** 2239–2299. MR2473656

[8] DAI, J. G. and PRABHAKAR, B. (2000). The throughput of switches with and without speed-up. In *Proceedings of IEEE Infocom* 556–564. IEEE, New York.

[9] DEMERS, A., KESHAV, S. and SHENKER, S. (1990). Analysis and simulation of a fair queuing algorithm. *Internetworking*: *Research and Experience* **1** 3–26.

[10] EL GAMAL, A., MAMMEN, J., PRABHAKAR, B. and SHAH, D. (2006). Optimal throughput-delay scaling in wireless networks. II. Constant-size packets. *IEEE Trans. Inform. Theory* **52** 5111–5116. MR2300378

[11] FOSS, S. and KONSTANTOPOULOS, T. (2004). An overview of some stochastic stability methods. *J. Oper. Res. Soc. Japan* **47** 275–303. MR2174067

[12] GALLAGER, R. and PAREKH, A. (1993). A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Transactions on Networking* **1** 344–357.

[13] GANESH, A., O'CONNELL, N. and WISCHIK, D. (2004). *Big Queues*. Springer, Berlin. MR2045489

[14] HARRISON, J. M. (1995). Balanced fluid models of multiclass queueing networks: A heavy traffic conjecture. In *Stochastic Networks*. *IMA Vol. Math. Appl*. **71** 1–20. Springer, New York. MR1381003

[15] HARRISON, J. M. (2000). Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probab*. **10** 75–103. MR1765204

[16] HARRISON, J. M. (2003). Correction: "Brownian models of open processing networks: Canonical representation of workload" [*Ann. Appl. Probab*. **10** (2000), no. 1, 75–103; MR1765204 (2001g:60230)]. *Ann. Appl. Probab*. **13** 390–393. MR1952004

[17] JAGABATHULA, S. and SHAH, D. (2008). Optimal delay scheduling in networks with arbitrary constraints. In *Proceedings of the* 2008 *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* 395–406. ACM, New York.

[18] KANG, W. N., KELLY, F. P., LEE, N. H. and WILLIAMS, R. J. (2009). State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probab.* **19** 1719–1780. MR2569806

[19] KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester. MR0554920

[20] KELLY, F. P., MASSOULIÉ, L. and WALTON, N. S. (2009). Resource pooling in congested networks: Proportional fairness and product form. *Queueing Syst.* **63** 165–194. MR2576010

[21] KINGMAN, J. F. C. (1962). On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B* **24** 383–392. MR0148146

[22] MATOUŠEK, J. (2002). *Lectures on Discrete Geometry*. Springer, New York. MR1899299

[23] MEYN, S. (2009). Stability and asymptotic optimality of generalized maxweight policies. *SIAM J. Control Optim.* **47** 3259–3294. MR2476439

[24] MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London. MR1287609

[25] NEELY, M., MODIANO, E. and CHENG, Y. (2007). Logarithmic delay for $n \times n$ packet switches under the crossbar constraint. *IEEE/ACM Transactions on Networking* (*TON*) **15** 657–668.

[26] PROUTIÈRE, A. (2003). Insensitivity and stochastic bounds in queuing netwokrs—applications to flow level traffic modelling in telecommunication networks. Ph.D. thesis, Ecole Doctorale, de l'Ecole Polytechnique.

[27] SHAH, D., TSITSIKLIS, J. N. and ZHONG, Y. Queue-size scaling for input-queued switches. Unpublished manuscript.

[28] SHAH, D., TSITSIKLIS, J. N. and ZHONG, Y. (2010). Qualitative properties of α-weighted scheduling policies. In *ACM SIGMETRICS Performance Evaluation Review* 239–250. ACM, New York.

[29] SHAH, D., TSITSIKLIS, J. N. and ZHONG, Y. (2011). Optimal scaling of average queue sizes in an input-queued switch: An open problem. *Queueing Syst.* **68** 375–384. MR2834209

[30] SHAH, D., TSITSIKLIS, J. N. and ZHONG, Y. (2014). Qualitative properties of α-fair policies in bandwidth-sharing networks. *Ann. Appl. Probab.* **24** 76–113. MR3161642

[31] SHAH, D. and WISCHIK, D. (2011). Fluid models of congestion collapse in overloaded switched networks. *Queueing Syst.* **69** 121–143. MR2836736

[32] SHAH, D. and WISCHIK, D. (2012). Switched networks with maximum weight policies: Fluid approximation and multiplicative state space collapse. *Ann. Appl. Probab.* **22** 70–127. MR2932543

[33] STOLYAR, A. L. (2004). Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.* **14** 1–53. MR2023015

[34] STOLYAR, A. L. (2008). Large deviations of queues sharing a randomly time-varying server. *Queueing Syst.* **59** 1–35. MR2429895

[35] TASSIULAS, L. and EPHREMIDES, A. (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automat. Control* **37** 1936–1948. MR1200609

[36] VENKATARAMANAN, V. J. and LIN, X. (2007). Structural properties of LDP for queue-length based wireless scheduling algorithms. In 45*th Annual Allerton Conference on Communication*, *Control*, *and Computing* 759–766. IEEE, New York.

[37] VON NEUMANN, J. (1953). A certain zero-sum two-person game equivalent to the optimal assignment problem. In *Contributions to the Theory of Games*. 5–12. Princeton Univ. Press, Princeton, NJ. MR0054920

[38] WALTON, N. S. (2009). Proportional fairness and its relationship with multi-class queueing networks. *Ann. Appl. Probab.* **19** 2301–2333. MR2588246

[39] WHITT, W. (2002). *Stochastic-Process Limits*. Springer, New York. MR1876437

[40] WILLIAMS, R. J. (1998). Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Syst.* **30** 27–88. MR1663759

[41] ZACHARY, S. (2007). A note on insensitivity in stochastic networks. *J. Appl. Probab.* **44** 238–248. MR2312999

D. SHAH
DEPARTMENT OF EECS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: devavrat@mit.edu

N. WALTON
KORTEWEG-DE VRIES INSTITUTE FOR MATHEMATICS
UNIVERSITY OF AMSTERDAM
1090 GE AMSTERDAM
THE NETHERLANDS
E-MAIL: n.s.walton@uva.nl

Y. ZHONG
DEPARTMENT OF IEOR
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: yz2561@columbia.edu