Optimal Queue-Size Scaling in Switched Networks

Devavrat Shah
MIT, LIDS
Cambridge, MA 02139, USA
devavrat@mit.edu

Neil Walton
University of Amsterdam
1012 ZA Amsterdam
The Netherlands
n.s.walton@uva.nl

Yuan Zhong
MIT, LIDS
Cambridge, MA 02139, USA
zhyu4118@mit.edu

ABSTRACT

We consider a switched (queueing) 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 average queue-size scaling for a class of switched networks including input-queued switches. In particular, it establishes the validity of a conjecture (documented in [24]) about optimal queue-size scaling for input-queued switches.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Packet Switching Networks; G.3 [**Probability and Statistics**]: Markov Processes, Queueing Theory, Stochastic Processes

General Terms

Algorithms, Performance, Theory

Keywords

Switched Network, Markov Chain, Large Deviations, Emulation, Heavy Traffic, Store-and-Forward

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, i.e., the network is single-hop.

Switched networks are special cases of what Harrison [12] 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 [31]. Finally, they can be instrumental in finding the right operational point in a data center [27].

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 average queue size scales[1] 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 queueing 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 non-empty queues. Then, the average total queue size scales as $1/(1-\rho)$. Indeed, these are instances of switched networks that differ in their scheduling

[1]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.

SIGMETRICS'12, June 11–15, 2012, London, England, UK.

set $\mathcal{S}$, which leads to different queue-size scalings. Therefore, a natural question is the determination of average 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, $\sqrt{N}/(1-\rho)$ [24].

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 the conjecture stated in [24] 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 *any* single-hop switched network.

We note that the validity of the conjecture in [24] for input-queued switches, stating that optimal average total queue size scales as $\sqrt{N}/(1-\rho)$, is a significant improvement over the best known bounds of $O(N/(1-\rho))$ (due to the moment bounds of [20] for the maximum weight policy) or $O\left(\frac{\sqrt{N}\log N}{(1-\rho)^2}\right)$ (obtained by using a batching policy [21]).

Our analysis consists of two principal components. Firstly, 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 queueing 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 queueing process becomes too large then an appropriate schedule is allocated. Secondly, we implement specific bandwidth allocation named the store-and-forward allocation policy (SFA). This policy was first considered by Massoulié (see page 63, sec 3.4.1 [22]), and was consequently discussed in Section 3.4 of Proutière's thesis [22]. 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 [37]. The Store-and-Forward bandwidth allocation policy is closely related to classical product-form multiclass queueing network, which have highly desirable queue-size scalings. By emulating these queueing networks, we are able to translate results which render optimal queue-size bounds for a switched network.

We make two remarks here. First, the focus of this paper is on policies that achieve provably (close-to-)optimal performance bounds, but not on their implementation complexity. While the design of low-complexity, and hence practically implementable, scheduling policies with provably optimal performance remains an important challenge, we believe that our result is a substantial advancement in this direction. Second, besides classical product-form queueing networks, SFA is also closely related to the so-called proportionally fair bandwidth allocation (see, for example, [16]), and their relationship is explored in [34]. Proportional fairness can be viewed as a low-complexity approximation to SFA, and it is natural to conjecture that a version of proportional fairness is optimal for a large class of switched networks, including the input-queued switch. See Section 3 for more discussions.

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 $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ be the set of non-negative integers, $\mathbb{R}$ the set of real numbers, and $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geqslant 0\}$.

We will reserve bold letters for vectors in $\mathbb{R}^N$, where N is the number of queues. For example, $\mathbf{x} = [x_n]_{1 \leqslant n \leqslant N}$. 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 $\mathbf{1}$ be 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 $\mathbf{v}$, and functions $f : \mathbb{R} \to \mathbb{R}$, we let $\mathbf{u}\cdot\mathbf{v} = \sum_{n=1}^N u_n v_n$, $\mathbf{uv} = [u_n v_n]_{1 \leqslant n \leqslant N}$, and $f(\mathbf{u}) = \big[f(u_n)\big]_{1 \leqslant n \leqslant N}$, and let matrix multiplication take precedence over dot product so that

$$\mathbf{u}\cdot\mathbf{A}\mathbf{v} = \mathbf{u}\cdot(\mathbf{A}\mathbf{v})\,.$$

For a set $\mathcal{S} \subset \mathbb{R}^N$, denote its convex hull by $\langle \mathcal{S} \rangle$.

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 $\tau \in \{0, 1, \dots\}$. Let $Q_i(\tau)$ be the amount of work in queue $i \in \{1, \dots, N\}$ at time slot τ. Following our general notation for vectors, we write $\mathbf{Q}(\tau)$ for $[Q_i(\tau)]_{1 \leqslant i \leqslant N}$. 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 queueing 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

$$\mathbf{Q}(\tau+1) = \big[\mathbf{Q}(\tau) - d\mathbf{B}(\tau)\big]^+ + d\mathbf{A}(\tau) \tag{1}$$

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

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

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

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

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

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

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

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

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, i.e., the cumulative arrival process $\mathbf{A}(\cdot)$ till time τ.

2.2 Stochastic model.

We shall assume that the exogeneous 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 $\boldsymbol{\sigma}' \leqslant \boldsymbol{\sigma}$ component-wise, 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 $\mathcal{S}$, i.e.,*

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

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

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

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 $J \geqslant 1$, a matrix $\mathbf{R} \in \mathbb{R}_+^{J\times N}$, and a vector $\mathbf{C} \in \mathbb{R}_+^J$ such that

$$\langle\mathcal{S}\rangle = \Big\{ \mathbf{x} \in [0,1]^N : \mathbf{R}\mathbf{x} \leqslant \mathbf{C} \Big\}. \tag{6}$$

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*

$$\text{minimize} \quad \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}} \tag{7}$$

$$\text{subject to} \quad \boldsymbol{\lambda} \leqslant \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}}\boldsymbol{\sigma}, \tag{8}$$

$$\alpha_{\boldsymbol{\sigma}} \in \mathbb{R}_+, \text{ for all } \boldsymbol{\sigma} \in \mathcal{S}. \tag{9}$$

Define the load induced 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 $\rho(\boldsymbol{\lambda}) \leqslant 1$. It also follows immediately from Definition 2.3 that

$$\rho(\boldsymbol{\lambda}) = \inf\Big\{\gamma \geqslant 0 : \mathbf{R}\boldsymbol{\lambda} \leqslant \gamma\mathbf{C}\Big\},$$

and $\boldsymbol{\lambda}$ is admissible if and only if $\mathbf{R}\boldsymbol{\lambda} \leqslant \mathbf{C}$, component-wise.

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

$$\rho(\mathbf{a}+\mathbf{b}) \leqslant \rho(\mathbf{a}) + \rho(\mathbf{b}). \tag{10}$$

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 $1 \leqslant i \leqslant N$.)

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 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 $\boldsymbol{\sigma} \in \mathbb{R}_+^{3\times3}$ 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 $\sum_{m=1}^3 \sigma_{k,m} \leqslant 1$ for $k = 1,2,3$, and $\sum_{m=1}^3 \sigma_{m,\ell} \leqslant 1$ 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 to be

$$\Big\{\boldsymbol{\sigma} \in \{0,1\}^{n\times n} : \sum_{m=1}^n \sigma_{k,m} \leqslant 1, \sum_{m=1}^n \sigma_{m,\ell} \leqslant 1, 1 \leqslant k,\ell \leqslant n\Big\}.$$

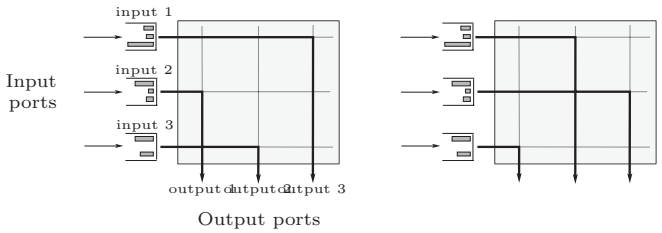

Figure 1: An input-queued switch, and two example matchings of inputs to outputs.

It can be checked that $\mathcal{S}$ is *monotone*. Furthermore, due to Birkhoff-von Neumann Theorem, [2, 33], the convex hull $\langle\mathcal{S}\rangle$ of $\mathcal{S}$ is given by

$$\Big\{\mathbf{x}\in[0,1]^{n\times n}:\sum_{m=1}^{n}x_{k,m}\leqslant 1,\sum_{m=1}^{n}x_{m,\ell}\leqslant 1, 1\leqslant k,\ell\leqslant n\Big\}.$$

That is, 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[2] $\boldsymbol{\lambda}\in[0,1]^{n\times n}$, $\rho(\boldsymbol{\lambda})$ is given by

$$\rho(\boldsymbol{\lambda})=\max_{1\leqslant k,\ell\leqslant n}\Big\{\sum_{m=1}^{n}\lambda_{k,m},\sum_{m=1}^{n}\lambda_{m,\ell}\Big\}.$$

3. RELATED WORKS.

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

In heavy-traffic analysis, one studies the queue-size behavior under a diffusion (or heavy-traffic) scaling. This regime was first considered by Kingman [18]; since then, a substantial body of theory has developed, and modern treatments can be found in [11, 5, 36, 35]. Stolyar [30] has studied a class of myopic scheduling policies, known as the maximum weight policy, introduced by Tassiulas and Ephremides [31], 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 [30], 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 $\alpha>0$. This condition effectively requires that there exists a scheduling policy which is able to balance the weights of all the heavily loaded queues. Stolyar [30] 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 [30, 6, 7] 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 approriately 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 [28]. 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 [27] 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 [19], 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. [15] have established a diffusion approximation for the proportionally fair bandwidth allocation policy, assuming a technical "local traffic" condition, but without assuming complete resource pooling[3]. They show that the resulting diffusion approximation has a product-form stationary distribution. Shah et al. [25] 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

[2]not a vector, for notational convenience, as discussed earlier

[3]Kang et al. [15] assume that critically loaded traffic is such that all the constraints are saturated simultaneously.

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. [15], building upon the work of Bramson [5] and Williams [36], 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 non-asymptotic, in the sense that they hold for any admissible load, they have a product-form structure, and they 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'.

We note that Stolyar [29] and Venkataramanan and Lin [32] 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. However, it does not characterize the tail exponent explicitly. See [23] 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 non-asymptotic 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 on 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 [21]. (Here N is the number of queues, and ρ the system load.) For more details of these results, see [24].

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 any single-hop switched network.

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. Justifications of claims made in this section are omitted in the interest of space.

We consider a bandwidth-sharing network operating in continuous time with capacity constraints. The particular bandwidth-sharing policy of interest is the so-called "store-and-forward allocation (SFA)," introduced by Bonald and Proutière [4]. 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 what we know about its performance.

Model.

Let time be continuous and indexed by $t \in \mathbb{R}_+$. Consider a network with $J \geqslant 1$ resources indexed from $1, \ldots, J$. Let there be N routes, and suppose that each *packet* on route i consumes an amount $R_{ji} \geqslant 0$ 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, i.e., $\mathcal{K} = \{(j, i) : R_{ji} > 0\}$. 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. 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

$$\mathbf{R}\boldsymbol{\phi}(\mathbf{m}) \leqslant \mathbf{C}, \quad \text{component-wise}, \tag{11}$$

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 $\mathbf{X}(t)$ 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 $\mathbf{X}(\cdot)$ to be positive (Harris) recurrent, it is necessary that

$$\mathbf{R}\boldsymbol{\lambda} \quad < \quad \mathbf{C}, \quad \text{component-wise}. \tag{12}$$

All such $\boldsymbol{\lambda} = [\lambda_i]_{i=1}^N \in \mathbb{R}_+^N$ will be called *strictly admissible*, in the same spirit as the admissible region for a switched network.

Store-and-Forward Allocation (SFA) policy.

We describe the store-and-forward allocation policy that was first considered by Massoulié and later analysed in the thesis of Proutière [22]. Bonald and Proutière [4] established that it induces product-form stationary distributions and is insensitive with respect to phase-type service distributions. This policy is shown to be *insensitive* for general service time

distributions, including the deterministic service considered here, by Zachary [37]. The relation between this policy, the proportionally fair allocation, and multiclass queueing networks is discussed in depth by Walton [34] and Kelly et al. [16]. 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 queueing networks. Consider a continuous-time multi-class queueing network (without scheduling constraints) consisting of processor sharing queues indexed by $j \in \{1, ..., J\}$ and job types indexed by the routes $i \in \{1, ..., N\}$. 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 jobs on each route arrive as a Poisson process, then the resulting queueing 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 [17]. 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 they do not obey the necessary scheduling constraints (11). However, these networks do emit an admissible schedule on average. For this reason, we consider SFA which, given the number of jobs on each route, allocates the average rate that jobs are transferred through this multi-class network. Next, we describe this policy (using notations similar to those used in [16, 34]).

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

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

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

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

Here, by notation $j \in i$ (and $i \ni j$) we mean $R_{ji} > 0$. The notation $i \ni j$ is used when we consider a collection of i satisfying this condition for a given j. For each $\widetilde{\mathbf{m}} \in U(\mathbf{m})$, we exploit notation somewhat and define

$$\widetilde{m}_j = \sum_{i:j\in i} \widetilde{m}_{ji}, \text{ for all } j \leqslant J.$$

Also define

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

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

$$\Phi(\mathbf{m}) = \sum_{\widetilde{\mathbf{m}}\in U(\mathbf{m})} \prod_{j\in J} \left(\binom{\widetilde{m}_j}{\widetilde{m}_{ji} : i \ni j} \prod_{i:j\in i} \left(\frac{R_{ji}}{C_j} \right)^{\widetilde{m}_{ji}} \right). \quad (13)$$

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 $\boldsymbol{\phi} : \mathbb{Z}_+^N \to \mathbb{R}_+^N$ so that for any $\mathbf{m} \in \mathbb{Z}_+^N$, $\boldsymbol{\phi}(\mathbf{m}) = (\boldsymbol{\phi}_i(\mathbf{m}))_{i=1}^N$, with

$$\boldsymbol{\phi}_i(\mathbf{m}) = \frac{\Phi(\mathbf{m} - \mathbf{e}_i)}{\Phi(\mathbf{m})}, \quad (14)$$

where, recall 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 $\phi(m)$ is the stationary throughput of jobs on the routes of a multi-class queueing 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 (i.e., satisfies (11)). 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 [16, Corollary 2] and [34, 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 $\mathbf{X}(t)$ is positive (Harris) recurrent and $\mathbf{M}(t)$ has a unique stationary probability distribution $\boldsymbol{\pi}$ given by*

$$\boldsymbol{\pi}(\mathbf{m}) = \frac{\Phi(\mathbf{m})}{\Phi} \prod_{i=1}^N \lambda_i^{m_i}, \quad \textit{for all } \mathbf{m} \in \mathbb{Z}_+^N, \quad (15)$$

where

$$\Phi = \prod_{j=1}^J \left(\frac{C_j}{C_j - \sum_{i:i\ni j} R_{ji}\lambda_i} \right) \quad (16)$$

is a normalizing factor. Furthermore, in steady state, the residual workload of each packet in the network is uniformly distributed on $[0,1]$ 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 statements similar to Theorem 4.1 have appeared in other works, for example, [3], [34, Proposition 4.2] and [16]. Theorem 4.1 is a summary of these statements.

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

PROPOSITION 4.2. *Consider the setup of Theorem 4.1 and let $\boldsymbol{\pi}$ be as described by (15). Define a measure $\widetilde{\boldsymbol{\pi}}$ on $\mathbb{Z}_+^{|\mathcal{K}|}$ as follows: for $\widetilde{\mathbf{m}} \in \mathbb{Z}_+^{|\mathcal{K}|}$,*

$$\widetilde{\boldsymbol{\pi}}(\widetilde{\mathbf{m}}) = \frac{1}{\Phi} \prod_{j=1}^J \left(\binom{\widetilde{m}_j}{\widetilde{m}_{ji} : i \ni j} \prod_{i:j\in i} \left(\frac{R_{ji}\lambda_i}{C_j} \right)^{\widetilde{m}_{ji}} \right). \quad (17)$$

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

$$\boldsymbol{\pi}\Big(\Big\{ \mathbf{m} : \sum_{i=1}^N m_i = L \Big\} \Big) = \widetilde{\boldsymbol{\pi}}\Big(\Big\{ \widetilde{\mathbf{m}} : \sum_{j=1}^J \widetilde{m}_j = L \Big\} \Big). \quad (18)$$

Finally, we relate the distribution $\widetilde{\boldsymbol{\pi}}$ to the stationary distribution of an insensitive multiclass queueing network with a product-form stationary distribution and geometrically distributed queue sizes.

PROPOSITION 4.3. *Consider the distribution $\widetilde{\boldsymbol{\pi}}$ defined in* (17). *Then, for any $\mathbf{L} = (L_1, \ldots, L_J) \in \mathbb{Z}_+^J$,*

$$\widetilde{\boldsymbol{\pi}}\big(\widetilde{m}_1 = L_1, \ldots, \widetilde{m}_J = L_J\big) \overset{(a)}{=} \sum_{(\tilde{m}_{ji}) \in V(\mathbf{L})} \widetilde{\boldsymbol{\pi}}((\widetilde{m}_{ji})) = \prod_{j=1}^{J} \widetilde{\rho}_j^{L_j}(1 - \widetilde{\rho}_j), \qquad (19)$$

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

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 any single-hop switched network 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))

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

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, the **BN** will operate under the insensitive SFA policy described in Section 4. Due to Theorem 4.1 as well as 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 policy SFA, 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 $\mathbf{A}(\cdot) = \big(A_i(\cdot)\big)$ 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 $\mathbf{M}(t) = \big(M_i(t)\big)$ denote the vector of numbers of packets waiting on the N routes in **BN** at time $t \geqslant 0$. In **BN**, the services are allocated according to the SFA policy described in Section 4. Let $\Lambda^{\mathrm{SFA}}(\cdot) = \big(\Lambda_i^{\mathrm{SFA}}(\cdot)\big) \in \mathbb{R}_+^N$ 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 $t \geqslant 0$, with $\Lambda_i^{\mathrm{SFA}}(0) = 0$ for $1 \leqslant i \leqslant N$. By definition, all components of $\Lambda^{\mathrm{SFA}}(\cdot)$ are non-decreasing and Lipschitz continuous. Furthermore, $(\Lambda^{\mathrm{SFA}}(t+s) - \Lambda^{\mathrm{SFA}}(t))/s \in \langle\mathcal{S}\rangle$ for any $t \geqslant 0$ 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 (14).

Now we describe the scheduling policy for **SN** that will rely on $\Lambda^{\mathrm{SFA}}(\cdot)$. Let $\mathbf{B}(\tau) = \big(B_i(\tau)\big)$ denote the cumulative amount of service allocated in **SN** by the scheduling policy up to time slot $\tau \geqslant 0$, with $\mathbf{B}(0) = \mathbf{0}$. The scheduling policy determines how $\mathbf{B}(\cdot)$ is updated. Let $\mathbf{Q}(\tau) = \big(Q_i(\tau)\big)$ be the queue sizes measured at the end of time slot τ. Let service be provided according to the scheduling policy instantly at the begining 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 by virtue of our policy, $\mathbf{D}(\tau) \geqslant d\mathbf{B}(\tau)$, and hence $\mathbf{D}(\tau) \geqslant \mathbf{0}$, for all time τ. This fact will be key for many subsequent proofs. 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 non-negative combination of schedules in $\mathcal{S}$ such that

$$\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}}\boldsymbol{\sigma} \geqslant \mathbf{D}(\tau), \quad \text{and} \quad \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \tilde{\alpha}_{\boldsymbol{\sigma}} = \rho(\mathbf{D}(\tau)). \qquad (20)$$

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

$$\sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}}\boldsymbol{\sigma} = \mathbf{D}(\tau), \quad \text{and} \quad \sum_{\boldsymbol{\sigma}\in\mathcal{S}} \alpha_{\boldsymbol{\sigma}} = \rho(\mathbf{D}(\tau)). \qquad (21)$$

This is formalized in the following lemma.

LEMMA 5.1. *Let $\mathbf{D} \in \mathbb{R}_+^N$ be a non-negative 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 exist $\alpha_{\boldsymbol{\sigma}} \geqslant 0$, $\boldsymbol{\sigma} \in \mathcal{S}$, such that* (21) *hold.*

The proof of the lemma relies on Assumption 2.1. The detail of the proof is provided in [26].

There could be many possible non-negative combinations of $\mathbf{D}(\tau)$ satisfying (21). If there exists non-negative numbers $\alpha_{\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} \in \mathcal{S}$, satisfying (21) with $\alpha_{\boldsymbol{\sigma}'} \geqslant 1$ 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

$$\text{maximize} \quad \sum_i \sigma_i \quad \text{over} \quad \boldsymbol{\sigma} \in \mathcal{S},\ \boldsymbol{\sigma} \leqslant \mathbf{D}(\tau). \qquad (22)$$

Note that $\mathbf{0}$ is a feasible solution for the above problem as $\mathbf{0} \in \mathcal{S}$ and $\mathbf{0} \leqslant \mathbf{D}(\tau)$. Observe also that for all time τ, $d\mathbf{B}(\tau) \leqslant \mathbf{D}(\tau)$.

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

THEOREM 5.2. *Given a strictly admissible arrival rate vector $\boldsymbol{\lambda}$, with $\rho(\boldsymbol{\lambda}) < 1$, under the above described policy, the switched network **SN** is positive recurrent and has a unique stationary distribution. Let $\tilde{\rho}_j = \left(\sum_i R_{ji}\lambda_i\right)/C_j$, $j = 1, 2, \ldots, J$ be as in Proposition 4.3. With respect to this stationary distribution, the following properties hold:*

1. *The expected total queue size is bounded as*

$$\mathbb{E}\Big[\sum_{i=1}^{N} Q_i\Big] \leqslant \frac{1}{2}\left(\sum_{j=1}^{J} \frac{\tilde{\rho}_j}{1-\tilde{\rho}_j}\right) + K(N+2), \qquad (23)$$

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

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

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{P}\Big(\sum_{i=1}^{N} Q_i \geqslant L\Big) = \max_{j=1,\ldots,J} \log \tilde{\rho}_j. \qquad (24)$$

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. The policy produces an optimal tail exponent for any single-hop switched network.

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 $\boldsymbol{\pi}$ 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 $\bar{Q}$ be the expected total queue size under the stationary distribution $\boldsymbol{\pi}$, defined by

$$\bar{Q} = \mathbb{E}_{\boldsymbol{\pi}}\Big[\sum_i Q_i\Big].$$

   Note that by ergodicity, the time average of the total queue size and the expected total queue size under $\boldsymbol{\pi}$ 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 $\boldsymbol{\pi}$ (possibly $-\infty$ or 0), respectively, defined by

$$\beta_L(Q) = \liminf_{\ell\to\infty} \frac{1}{\ell} \log \mathbb{P}_{\boldsymbol{\pi}}\Big(\sum_i Q_i \geqslant \ell\Big), \quad (25)$$

   and
$$\beta_U(Q) = \limsup_{\ell\to\infty} \frac{1}{\ell} \log \mathbb{P}_{\boldsymbol{\pi}}\Big(\sum_i Q_i \geqslant \ell\Big). \quad (26)$$

   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 $\bar{Q}$ and $\beta(Q)$. For tractability reasons, we focus on the *scaling* of these quantities with respect to $\mathcal{S}$ (equivalently, N) and $\rho(\boldsymbol{\lambda})$, as $1/(1-\rho(\boldsymbol{\lambda}))$ and N increase. Now, for different $\boldsymbol{\lambda}'$ and $\boldsymbol{\lambda}$, it is possible that $\rho(\boldsymbol{\lambda}) = \rho(\boldsymbol{\lambda}')$, but the scaling of $\bar{Q}$, for example, could be wildly different. For this reason, we consider the worst possible dependence on $1/(1-\rho)$ and N among all $\boldsymbol{\lambda}$ with $\rho(\boldsymbol{\lambda}) = \rho$.

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 average queue-size stated here is with respect to the order of limits $\rho \to 1$ and then $N \to \infty$. As noted in [24], taking the limits in different orders could potentially result in different limiting behaviors of the object of interest, e.g., $\bar{Q}$. For more discussions, see Section 6. It should be noted, however that the optimality of the tail exponent holds for *any* ρ and N.

Optimality of the tail exponent.

Here we establish the optimality of the tail exponent for any single-hop switched network under our policy. Consider any policy under which there exists a well-defined limiting stationary distribution of the queue sizes for all $\boldsymbol{\lambda}$ such that $\rho(\boldsymbol{\lambda}) < 1$. Let $\boldsymbol{\pi}_0$ denote the stationary distribution of queue sizes under this policy. The optimality of the tail exponent under our policy is an immediate consequence of the following lemma.

LEMMA 5.3. *Let $\boldsymbol{\pi}_0$ and $\boldsymbol{\lambda}$ be as described. Let $\tilde{\rho}_1, \ldots, \tilde{\rho}_J$ be as defined in* (4.3)*. Then under $\boldsymbol{\pi}_0$,*

$$\liminf_{\ell\to\infty} \frac{1}{\ell} \log \mathbb{P}_{\boldsymbol{\pi}_0}\left(\sum_i Q_i \geqslant \ell\right) = \max_{j=1,2,\ldots,J} \log \tilde{\rho}_j.$$

PROOF. Recall that $\tilde{\rho}_j = \left(\sum_i R_{ji}\lambda_i\right)/C_j$, for $j = 1, 2, \ldots, J$, under the representation

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

Without loss of generality, suppose that $\tilde{\rho}_1 = \max_{j=1,2,\ldots,J} \tilde{\rho}_j$. We now lower bound the total queue size stochastically by that of an $M/D/1$ queue. Consider an $M/D/1$ queue where the arrival rate is $\tilde{\rho}_1$, and the service capacity has a deterministic rate of 1. Since in the original network, this service capacity has to be shared among the queues, $\sum_i Q_i$ stochastically dominates this $M/D/1$ queue. Now the stationary distribution of this $M/D/1$ queue has a tail exponent $\log \tilde{\rho}_1$, which provides a lower bound on the same quantity in the original network, under $\boldsymbol{\pi}_0$. □

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 any switched networks, and hence 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 et al. [24], 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 5.4. *Consider a n-port input-queued switch, with an arrival rate vector $\boldsymbol{\lambda}$. Suppose that the loads on all input and output ports are ρ, i.e., $\sum_{k=1}^{n}\lambda_{k,\ell}=\sum_{m}\lambda_{\ell,m}=\rho$, 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*

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

PROOF. We consider the sums of queue sizes at each output port, i.e., 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 $\ell\in\{1,2,\ldots,n\}$,

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

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. □

The optimality in terms of the average total queue size is a direct consequence of Theorem 5.2 and Lemma 5.4.

COROLLARY 5.5. *Consider the same setup as in Lemma 5.4. Then in the heavy-traffic limit $\rho\to1$, 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, i.e., $(1-\rho)\bar{Q}$. Then*

$$\limsup_{\rho\to1}(1-\rho)\bar{Q}\leqslant n$$

under our policy, and

$$\liminf_{\rho\to1}(1-\rho)\bar{Q}\geqslant\frac{n}{2}$$

under any other policy.

PROOF. Lemma 5.4 implies that

$$\liminf_{\rho\to1}(1-\rho)\bar{Q}\geqslant\frac{n}{2}$$

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

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

For input-queued switches, $J\leqslant2n$, as remarked in Section 5.2, $N=n^2$, and $K=n$. Therefore, we have that under our policy, the expected total queue size scales as

$$\bar{Q}\leqslant\frac{n}{1-\rho}+(n^2+2)n. \tag{27}$$

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

$$(1-\rho)\bar{Q}\leqslant n+(1-\rho)(n^2+2)n. \tag{28}$$

The term $(1-\rho)(n^2+2)n$ goes to zero as $\rho\to1$, and hence under our policy,

$$\limsup_{\rho\to1}(1-\rho)\bar{Q}\leqslant n.$$

□

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

5.3 Proof of Theorem 5.2.

The proof is divided in three parts. The first part describes a sample-path-wise relation between $\mathbf{Q}(\cdot)$ and $\mathbf{M}(\cdot)$, which implies that $\mathbf{Q}(\cdot)$ is essentially dominated by $\mathbf{M}(\cdot)$ 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. Given the technical nature of the second part, we skip the detail, which can be found in [26]. 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 5.8, which is a consequence of Lemmas 5.6 and 5.7 below.

LEMMA 5.6. *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 $\tau\geqslant0$ and $1\leqslant i\leqslant N$,*

$$Q_i(\tau)\leqslant W_i(\tau)+D_i(\tau)\ \leqslant M_i(\tau)+D_i(\tau), \tag{29}$$

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

PROOF. Consider any $i\in\{1,2,\ldots,N\}$ and $\tau\geqslant0$. From (4), in **SN**,

$$Q_i(\tau)=A_i(\tau)-B_i(\tau)+Z_i(\tau), \tag{30}$$

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

$$W_i(\tau)=A_i(\tau)-\Lambda_i^{\mathrm{SFA}}(\tau)+\widehat{Z}_i(\tau), \tag{31}$$

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 $\mathbf{D}(\tau)\geqslant\mathbf{0}$, we have that

$$B_i(\tau)\leqslant\Lambda_i^{\mathrm{SFA}}(\tau)\ \leqslant B_i(\tau)+D_i(\tau). \tag{32}$$

By definition, the instantaneous rate allocation to the ith queue satisfies $\frac{d}{dt^+}\Lambda_i^{\mathrm{SFA}}(t)=0$ if $W_i(t)=0$ (equivalently, if $M_i(t)=0$) for any $t\geqslant0$. Therefore, $\widehat{Z}_i(\tau)=0$. On the other hand, by Skorohod's map,

$$\begin{aligned}Z_i(\tau)&=\sup_{0\leqslant s\leqslant\tau}\left[B_i(s)-A_i(s)\right]^+\\&\leqslant\sup_{0\leqslant s\leqslant\tau}\left[\Lambda_i^{\mathrm{SFA}}(s)-A_i(s)\right]^+\\&=\widehat{Z}_i(\tau).\end{aligned} \tag{33}$$

From (32) and (33), it follows that

$$\begin{aligned} Q_i(\tau) &= A_i(\tau) - B_i(\tau) + Z_i(\tau) \\ &\leqslant A_i(\tau) - \Lambda_i^{\mathrm{SFA}}(\tau) + D_i(\tau) + Z_i(\tau) \\ &\leqslant A_i(\tau) - \Lambda_i^{\mathrm{SFA}}(\tau) + D_i(\tau) + \widehat{Z}_i(\tau) \\ &= W_i(\tau) + D_i(\tau). \end{aligned} \tag{34}$$

Since the workload at the ith queue equals the total amount of unfinished work for all of the $M_i(\tau)$ packets waiting at the ith queue, and since each packet has at most a unit amout of unfinished work, $W_i(\tau) \leqslant M_i(\tau)$. $\square$

LEMMA 5.7. *Let $\mathbf{D}(\tau)$ be as in Lemma 5.6. For all $\tau \geqslant 0$, $\rho(\mathbf{D}(\tau)) \leqslant N+2$. In particular,*

$$\sum_i D_i(\tau) \leqslant K(N+2), \quad \textit{where} \quad K = \max_{\boldsymbol{\sigma}\in\mathcal{S}} \sum_i \sigma_i. \tag{35}$$

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 $\rho(\mathbf{D}(\tau+1)) \leqslant \rho(\mathbf{D}(\tau)) + 1$. That is, $\rho(\mathbf{D}(\cdot))$ can at most increase by 1 in each time slot. And finally, we show that it cannot increase once it exceeds $N+1$. That is, if $\rho(\mathbf{D}(\tau)) \geqslant N+1$, then $\rho(\mathbf{D}(\tau+1)) \leqslant \rho(\mathbf{D}(\tau))$. This will complete the proof.

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

$$\begin{aligned} \mathbf{D}(\tau+1) &= \Lambda^{\mathrm{SFA}}(\tau+1) - \mathbf{B}(\tau+1) \\ &= \Lambda^{\mathrm{SFA}}(\tau) - \mathbf{B}(\tau) \qquad (36) \\ &\quad + \Big(\Lambda^{\mathrm{SFA}}(\tau+1) - \Lambda^{\mathrm{SFA}}(\tau) - d\mathbf{B}(\tau)\Big) \\ &= \mathbf{D}(\tau) + d\Lambda^{\mathrm{SFA}}(\tau) - d\mathbf{B}(\tau) \\ &= \Big(\mathbf{D}(\tau) - d\mathbf{B}(\tau)\Big) + d\Lambda^{\mathrm{SFA}}(\tau), \qquad (37) \end{aligned}$$

where $d\Lambda^{\mathrm{SFA}}(\tau) = \Lambda^{\mathrm{SFA}}(\tau+1) - \Lambda^{\mathrm{SFA}}(\tau)$. As remarked earlier, $d\mathbf{B}(\tau) \leqslant \mathbf{D}(\tau)$ component-wise. Therefore, by (10) it follows that

$$\rho\big(\mathbf{D}(\tau+1)\big) \leqslant \rho\big(\mathbf{D}(\tau) - d\mathbf{B}(\tau)\big) + \rho(d\Lambda^{\mathrm{SFA}}(\tau)).$$

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

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

Hence it follows that

$$\rho\big(\mathbf{D}(\tau+1)\big) \leqslant \rho(\mathbf{D}(\tau)) + 1. \tag{38}$$

Next, we shall argue that if $\rho(\mathbf{D}(\tau)) \geqslant N+1$, then $\rho(\mathbf{D}(\tau+1)) \leqslant \rho(\mathbf{D}(\tau))$. To that end, suppose that $\rho(\mathbf{D}(\tau)) \geqslant N+1$. 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 $\alpha_k \geqslant 0$ 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

$$\frac{1}{\rho(\mathbf{D}(\tau))}\mathbf{D}(\tau) = \sum_{k=1}^{N+1} \alpha_k \boldsymbol{\sigma}^k. \tag{39}$$

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

$$\mathbf{D}(\tau) - {\boldsymbol{\sigma}^k}^* = \sum_{k=1,k\neq k^*}^{N+1} \rho(\mathbf{D}(\tau))\alpha_k \boldsymbol{\sigma}^k + (\rho(\mathbf{D}(\tau))\alpha_{k^*} - 1){\boldsymbol{\sigma}^k}^*. \tag{40}$$

Therefore,

$$\rho\Big(\mathbf{D}(\tau) - {\boldsymbol{\sigma}^k}^*\Big) \leqslant \rho(\mathbf{D}(\tau)) - 1. \tag{41}$$

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

$$\mathbf{D}(\tau+1) = \mathbf{D}(\tau) - {\boldsymbol{\sigma}^k}^* + d\Lambda^{\mathrm{SFA}}(\tau). \tag{42}$$

By another application of (10) it follows that

$$\begin{aligned} \rho(\mathbf{D}(\tau+1)) &\leqslant \rho\big(\mathbf{D}(\tau) - {\boldsymbol{\sigma}^k}^*\big) + \rho(d\Lambda^{\mathrm{SFA}}(\tau)) \\ &\leqslant \rho\big(\mathbf{D}(\tau)\big) - 1 + 1, \\ &= \rho\big(\mathbf{D}(\tau)\big), \end{aligned} \tag{43}$$

where again we have used the fact that $\rho(d\Lambda^{\mathrm{SFA}}(\tau)) \leqslant 1$, due to the feasibility of SFA policy and (41). This establishes that $\rho(\mathbf{D}(\tau)) \leqslant N+2$ for all $\tau \geqslant 0$. That is, for each $\tau \geqslant 0$, there exist $\alpha_{\boldsymbol{\sigma}} \geqslant 0$ for all $\boldsymbol{\sigma} \in \mathcal{S}$, $\sum_{\boldsymbol{\sigma}} \alpha_{\boldsymbol{\sigma}} \leqslant N+2$ and

$$\mathbf{D}(\tau) \leqslant \sum_{\boldsymbol{\sigma}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}. \tag{44}$$

Therefore,

$$\begin{aligned} \sum_i D_i(\tau) &= \mathbf{D}(\tau)\cdot\mathbf{1} \\ &\leqslant \sum_{\boldsymbol{\sigma}} \alpha_{\boldsymbol{\sigma}} \boldsymbol{\sigma}\cdot\mathbf{1} \\ &\leqslant \Big(\sum_{\boldsymbol{\sigma}} \alpha_{\boldsymbol{\sigma}}\Big)\Big(\max_{\boldsymbol{\sigma}\in\mathcal{S}} \sum_i \sigma_i\Big) \\ &\leqslant (N+2)K, \end{aligned} \tag{45}$$

where $K = \max_{\boldsymbol{\sigma}\in\mathcal{S}} \sum_i \sigma_i$. This completes the proof of Lemma 5.7. $\square$

Lemma 5.6 and 5.7 together imply the following proposition.

PROPOSITION 5.8. *Let $\mathbf{Q}(\cdot)$, $\mathbf{W}(\cdot)$ and $\mathbf{M}(\cdot)$ be as in Lemma 5.6. Then*

$$\sum_{i=1}^N Q_i(\tau) \leqslant \sum_{i=1}^N W_i(\tau) + K(N+2) \leqslant \sum_{i=1}^N M_i(\tau) + K(N+2), \tag{46}$$

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

PROOF. We obtain the bounds (46) by summing inequality (29) over $i \in \{1,2,\ldots,N\}$, and using the bound (35). $\square$

Part 2. Positive recurrence. See [26].

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

$\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{i=1}^{N} W_i\right]$, where, similar to Lemma 5.6, W_i is the steady-state workload on queue i in **BN**. Define $\overline{M}$ similarly. By ergodicity, the time average of the total queue size equals the expected total queue size in steady state, i.e., $\bar{Q}$, and similarly for $\overline{W}$. Therefore, by Proposition 5.8,

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

We now claim that

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

First, we have that

$$\overline{M} \leqslant \sum_{j=1}^{J} \frac{\widetilde{\rho}_j}{1-\widetilde{\rho}_j}.$$

By Propositions 4.2 and 4.3, $\overline{M}$ is the sum of J independent geometric random variables, with parameters $1-\widetilde{\rho}_1, 1-\widetilde{\rho}_2, \ldots, 1-\widetilde{\rho}_J$. Hence, in fact, we have

$$\overline{M} = \sum_{j=1}^{J} \frac{\widetilde{\rho}_j}{1-\widetilde{\rho}_j}.$$

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 $[0,1]$. Therefore, $\overline{W} = \frac{1}{2}\overline{M}$, and the claim is proved.

We now establish the tail exponent in (24). By Lemma 5.3,

$$\beta_L(\mathbf{Q}) \geqslant \max_{j=1,2,\ldots,J} \log \widetilde{\rho}_j,$$

so we only need to show that

$$\beta_U(\mathbf{Q}) \leqslant \max_{j=1,2,\ldots,J} \log \widetilde{\rho}_j,$$

where $\beta_L(\mathbf{Q})$ and $\beta_U(\mathbf{Q})$ are defined in (25) and (26) respectively.

First note that $\beta_U(\mathbf{Q}) \leqslant \beta_U(\mathbf{M})$, where $\mathbf{M}$ is the queue-size vector of the virtual system **BN**. This is because, by Proposition 5.8,

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

deterministically and for all times τ. Thus, in steady state, $\sum_{i=1}^{N} Q_i$ is upper bounded by $\sum_{i=1}^{N} M_i(\tau) + K(N+2)$, deterministically. Since $K(N+2)$ is a constant, $\sum_{i=1}^{N} M_i(\tau) + K(N+2)$ and $\sum_{i=1}^{N} M_i(\tau)$ have the same tail exponent. This establishes that $\beta_U(\mathbf{Q}) \leqslant \beta_U(\mathbf{M})$.

We now consider $\beta_U(\mathbf{M})$. As noted earlier, in steady state, $\mathbf{M}$ is the sum of J independent geometric random variables, with parameters $1-\widetilde{\rho}_1, 1-\widetilde{\rho}_2, \ldots, 1-\widetilde{\rho}_J$. The following lemma states that the tail exponent of the sum of these J geometric random variables is upper bounded by $\max_{j=1,2,\ldots,J} \log \widetilde{\rho}_j$.

LEMMA 5.9. *Let M be the sum of J independent geometric random variables, with parameters $1-\widetilde{\rho}_1, \ldots, 1-\widetilde{\rho}_J$ respectively, where $\widetilde{\rho}_j \in [0,1)$ for all $j \in \{1,2,\ldots,J\}$. Then we have*

$$\limsup_{\ell\to\infty} \frac{1}{\ell} \log \mathbb{P}\left(M \geqslant \ell\right) \leqslant \max_{j=1,2,\ldots,J} \log \widetilde{\rho}_j.$$

The detailed proof is provided in [26].

In conclusion,

$$\beta_U(\mathbf{Q}) \leqslant \beta_U(\mathbf{M}) = \max_{j=1,2,\ldots,J} \log \widetilde{\rho}_j.$$

6. DISCUSSION.

We presented a novel scheduling policy for a generic 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 multi-class queueing 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 any single-hop switched network 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 [24]. 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 on its own may be of potential interest in other domains of applications.

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 [9] and [14]: all arriving packets are first passed through a 'regularizer', which emits 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, i.e., $(1-\rho)$, is decreased by factor 2.

The scheduling policy that we propose is not optimal for general switched networks. For example, in the context of ad-hoc wireless networks, in the independent-set model, 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 policies, which has polynomial complexity, but their optimality appears difficult to analyze. Another possible candidate could be, as discussed in the introduction, a (randomized) version of proportional fairness.

Acknowledgements.

Devavrat Shah and Yuan Zhong would like to thank John Tsitsiklis for a careful reading of the paper which has helped improve the readability, for his insights and support of this project. They would also like to acknowledge the support of NSF TF collaborative project and NSF CNS CAREER project.

7. REFERENCES

[1] S. Asmussen. *Applied probability and queues*. Second edition, Springer Verlag, 2003.

[2] G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman Rev. Ser. A*, 5:147–151, 1946.

[3] T. Bonald and A. Proutiére. Insensitivity in processor-sharing networks. *Performance Evaluation*, 49(1-4):193–209, 2002.

[4] T. Bonald and A. Proutiére. Insensitive bandwidth sharing in data networks. *Queueing systems*, 44(1):69–100, 2003.

[5] M. Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, 30:89–148, 1998.

[6] J. G. Dai and W. Lin. Maximum pressure policies in stochastic processing networks. *Operations Research*, 53(2), 2005.

[7] J. G. Dai and W. Lin. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *The Annals of Applied Probability*, 18(6), 2008.

[8] J. G. Dai and B. Prabhakar. The throughput of switches with and without speed-up. In *Proceedings of IEEE Infocom*, pages 556–564, 2000.

[9] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah. Optimal throughput-delay scaling in wireless networks – part ii: Constant-size packets. *Information Theory, IEEE Transactions on*, 52(11):5111–5116, 2006.

[10] S. Foss and T. Konstantopoulos. An overview of some stochastic stability methods. *Journal of Operations Research, Society of Japan*, 47(4), 2004.

[11] J. M. Harrison. Balanced fluid models of multiclass queueing networks: A heavy traffic conjecture. *Stochastic Networks*, 71:1–20, 1995. IMA Volumes in Mathematics and Its Applications.

[12] J. M. Harrison. Brownian models of open processing networks: canonical representation of workload. *The Annals of Applied Probability*, 10:75–103, 2000. Also see [13].

[13] J. M. Harrison. Correction to [12]. *The Annals of Applied Probability*, 13:390–393, 2003.

[14] S. Jagabathula and D. Shah. Optimal delay scheduling in networks with arbitrary constraints. In *Proceedings of the 2008 ACM SIGMETRICS*, pages 395–406. ACM, 2008.

[15] W. Kang, F. Kelly, N. Lee, and R. Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *The Annals of Applied Probability*, 2009.

[16] F. Kelly, L. Massoulié, and N. Walton. Resource pooling in congested networks: proportional fairness and product form. *Queueing Systems*, 63(1):165–194, 2009.

[17] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chicester, 1979.

[18] J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society, series B*, 24(2):383–392, 1962.

[19] S. Meyn. Stability and asymptotic optimality of generalized maxweight policies. *SIAM J. Control and Optimization*, 2008.

[20] S. Meyn and R. Tweedie. *Markov chains and stochastic stability*. Springer New York, 1993.

[21] M. Neely, E. Modiano, and Y. Cheng. Logarithmic delay for n× n packet switches under the crossbar constraint. *IEEE/ACM Transactions on Networking (TON)*, 15(3):657–668, 2007.

[22] A. Proutiére. *Insensitivity and stochastic bounds in queueing networks–Applications to flow level traffic modelling in telecommunication networks*. PhD thesis, Ecoloe Doctorale de l'Ecole Polytechnique, 2003.

[23] D. Shah, J. N. Tsitsiklis, and Y. Zhong. Qualitative properties of α-weighted scheduling policies. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 239–250. ACM, 2010.

[24] D. Shah, J. N. Tsitsiklis, and Y. Zhong. Optimal scaling of average queue sizes in an input-queued switch: an open problem. *Queueing Systems*, 68(3-4):375–384, 2011.

[25] D. Shah, J. N. Tsitsiklis, and Y. Zhong. Qualitative properties of alpha-fair policies in bandwidth sharing network. *Unpublished, available on arxiv.org*, 2011.

[26] D. Shah, N. Walton, and Y. Zhong. Optimal queue-size scaling in switched networks. *http://arxiv.org/pdf/1110.4697v1.pdf*.

[27] D. Shah and D. Wischik. Fluid models of congestion collapse in overloaded switched networks. *Queueing Systems*, 69(2):121–143, 2011.

[28] D. Shah and D. Wischik. Switched networks with maximum weight policies: Fluid approximation and state space collapse. *The Annals of Applied Probability*, 2011.

[29] A. Stolyar. Large deviations of queues sharing a randomly time-varying server. *Queueing Systems*, 59(1):1–35, 2008.

[30] A. L. Stolyar. MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 2004.

[31] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37:1936–1948, 1992.

[32] V. Venkataramanan and X. Lin. Structural properties of LDP for queue-length based wireless scheduling algorithms. In *Proceedings of Allerton*, 2007.

[33] J. von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *In Contributions to the theory of games*, 2, 1953.

[34] N. Walton. Proportional fairness and its relationship with multi-class queueing networks. *The Annals of Applied Probability*, 19(6):2301–2333, 2009.

[35] W. Whitt. *Stochastic-Process Limits*. Springer, 2001.

[36] R. J. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems*, 30:27–88, 1998.

[37] S. Zachary. A note on insensitivity in stochastic networks. *Journal of applied probability*, 44(1):238–248, 2007.