Fluid models of congestion collapse in overloaded switched networks

Citation: Shah, Devavrat, and Damon Wischik. "Fluid Models of Congestion Collapse in Overloaded Switched Networks." Queueing Systems 69.2 (2011): 121–143.

As Published: http://dx.doi.org/10.1007/s11134-011-9250-1

Publisher: Springer-Verlag

Persistent URL: http://hdl.handle.net/1721.1/73531

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Noname manuscript No.
(will be inserted by the editor)

Fluid models of congestion collapse in overloaded switched networks

Devavrat Shah · Damon Wischik

the date of receipt and acceptance should be inserted later

Abstract We consider a switched network (i.e. a queueing network in which there are constraints on which queues may be served simultaneously), in a state of overload. We analyse the behaviour of two scheduling algorithms for multihop switched networks: a generalized version of max-weight, and the α-fair policy. We show that queue sizes grow linearly with time, under either algorithm, and we characterize the growth rates. We use this characterization to demonstrate examples of congestion collapse, i.e. cases in which throughput drops as the switched network becomes more overloaded. We further show that the loss of throughput can be made arbitrarily small by the max-weight algorithm with weight function $f(q) = q^\alpha$ as $\alpha \to 0$.

Keywords fluid model, switch, bandwidth sharing, max-weight, overload

1 Introduction

By *switched network* we mean a collection of queues with restrictions on which queues may be served simultaneously. For example, consider a collection of four queues 1, 2, 3 and 4, operating in slotted time, and suppose that in each timeslot it is required that the offered service be either "Serve 1 unit of work each from queues 1 and 2" or "Serve 1 unit of work each from queues 2, 3 and 4". Switched networks have been used to model input-queued switches in Internet routers [5], multihop wireless networks [16], and bandwidth-sharing by the Internet's congestion control mechanism [1], and they can be used to model processing of jobs in data centres. Our general model is specified in Section 2, and some of these applications are described further in Section 4.

Devavrat Shah
Dept of EECS, MIT, Cambridge MA 02139
Tel. +1 617 253 4670, E-mail: devavrat@mit.edu

Damon Wischik
Dept of Computer Science, UCL, Gower St, London WC1E 6BT
Tel. +44 20 7679 0442, E-mail: d.wischik@cs.ucl.ac.uk

In this paper we analyse the behaviour of a switched network in overload, that is, when new work arrives at such a rate that queue sizes cannot be prevented from growing over time. A networked system has a certain capacity, and it is subject to varying load, and it may sometimes become overloaded—either because it is too expensive to build so much capacity that overload never occurs, or because it is too hard to predict at design time what future load will look like. For example, a web data centre may be underloaded most days, and show very good performance, but sometimes it may face a surge in demand which overloads the system. We would like 'graceful failure' under overload. For example, a classic $M/M/1$ queue fails gracefully, in the sense that throughput is always as high as it can be (namely, the minimum of the arrival rate and the service rate), regardless of the arrival rate. Engineers are however aware of many systems which exhibit 'congestion collapse', meaning that when arrival rates increase beyond a certain threshold, throughput actually drops[1]. In Section 4 we illustrate congestion collapse for some of the applications mentioned above.

The policy for deciding which queues get service, in a switched network, is called the scheduling policy. A natural question is 'Are there scheduling policies that avoid the problem of congestion collapse?' Our analysis in this paper is of two scheduling policies: a generalization of the max-weight policy of Tassiulas and Ephremides [16] (also known as the backpressure policy), and the α-fair policy described by Mo and Walrand [13] with $\alpha \geq 1$, in both single-hop and multihop switched networks. The α-fair policy has previously only been described for single-hop networks; our extension to multihop is novel. These policies are specified in Section 2. We show in Section 4.4 that the max-weight policy can nearly prevent congestion collapse, in the sense that it achieves throughput that is arbitrarily close to optimal, regardless of load. This is achieved by weighting the queues using the weight function $f(q) = q^\alpha$ for α sufficiently close to 0. It is remarkable that a single policy is near optimal for all switched networks of the general type considered in this paper (though the choice of α does depend on the number of queues).

Formally, our analysis is within the framework of fluid modelling. In this framework, one starts with a stochastic queueing model, and obtains limiting dynamics under a fluid scaling. The 'fluid model' describes these limiting dynamics. One then proves properties of the fluid model. Finally, results about the fluid model can be straightforwardly translated back into statements about rate-level behaviour of the original stochastic system. The focus of this paper is exclusively on analysis of the fluid model. The relationship between the stochastic model and the fluid model has already been established by Kelly and Williams [11] and Gromoll and Williams [8] for the bandwidth-sharing model of Internet congestion control running α-fair scheduling, and by Dai and Prabhakar [5] and others [3, 4, 15] for queueing networks running max-weight. We will briefly describe some stochastic models in Section 4, and their fluid limits, but for all the rest of the paper we will work exclusively with fluid models.

The main technical result of this paper is the following. If $\mathbf{q}(t)$ is the vector of queue sizes at time t, under a fluid model, then $\mathbf{q}(t)/t \to \hat{\mathbf{q}}$ as $t \to \infty$ for a particular vector $\hat{\mathbf{q}}$; if the queues start empty then $\mathbf{q}(t)/t = \hat{\mathbf{q}}$ for all $t > 0$. We identify $\hat{\mathbf{q}}$ as the solution to a certain optimization problem. Technically this result holds for all arrival rates, not just overload, but $\hat{\mathbf{q}}$ is only non-zero in overload; if the system is underloaded the result implies

[1] By 'congestion collapse' we mean some loss of throughput—this is how the term is used for example in [12]. We do not necessarily mean a near-complete loss of throughput, as in the Internet's 1986 congestion collapse [10].

weak stability, cf. [5]. We state this main result in Section 3, and the proofs are given in sections 5 and 6 for the maxweight and α-fair policies respectively. A distinctive feature of our presentation is that we work with a single unified model for both policies, and the proof of the main theorem is the same for both policies. However, the proof is based on Lyapunov techniques, and the Lyapunov functions are different for the two policies, which is why we divide the proofs into two sections.

1.1 Related work

Much of the theoretical work on switched networks, starting with [16], has studied stability. The stability region for a switched network is the set of arrival rates for which an omniscient scheduler can keep the network stable. An algorithm which is stable for all arrival rates in the interior of the stability region is said to have 100% throughput. The focus of much theoretical work has been on finding scheduling algorithms with 100% throughput.

Recently there have been attempts to understand the behaviour of overloaded systems, i.e. systems where the arrival rates lie outside the stability region.

Harrison and Zeevi [9] studied staffing levels in call centers in the presence of varying demand. In their model, the arrival rate of calls each day is random, and the manager has to decide on staffing levels before that day's arrival rate is revealed. Then the day's actual arrival rate is revealed, and the manager finds out if the call center is overloaded or underloaded; he/she then solves a linear programming problem, in which the objective is to maximize revenue, the input data is the day's arrival rate, and the optimization variables are the fraction of effort each staff member devotes to each type of call. This optimization is called the 'static planning problem'.

Georgiadis and Tassiulas [7] investigated overload in a sensor network, modelled as a single-commodity flow problem. They take the arrival rate as fixed and given. They pose a static planning problem with the objective of maximizing throughput, and another static planning problem with the objective of maximizing the time until one of the queues fills its buffer, or equivalently of finding the most balanced set of queue growth rates. They show that for their specific network model, the two objectives can be met simultaneously. They also describe an iterative distributed algorithm, inspired by max-weight, for computing the solution to the static planning problem.

Egorova et al [6] studied overload in a model for bandwidth sharing in the Internet. In this model, the scheduling decisions are made 'myopically' (by the α-fair scheduling algorithm) without any knowledge of the arrival rates—unlike the static planning problem, in which the arrival rates are treated as known input variables. The appeal of myopic algorithms is that (one hopes) they respond well to fluctuations in arrival rates over a range of timescales. The authors show that $\mathbf{q}(t) = t\hat{\mathbf{q}}$ is a feasible fluid model solution, where $\hat{\mathbf{q}}$ is the solution to a certain optimization problem. Further description of their model and results, and how it relates to our work, is given in Section 4.5. What is fascinating about this result is that the scheduling algorithm can be thought of as 'implying' a certain optimization problem, in contrast to the earlier work which took the static planning problem as its starting point.

Subsequent to our work, Chan et al [2] have studied overload in a single-hop switched network running max-weight with weight function $f(q) = q$ and a constant vector of queue weights. They prove a special case of our main result, but they use direct techniques whereas

we use Lyapunov functions. They investigate how the queue weights may be chosen, in order to achieve a balanced set of queue growth rates.

1.2 Notation

Let $\mathbb{Z} = \{0, 1, \dots, \}$, let $\mathbb{R}$ be the set of real numbers, let $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$, and let $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$. Let N be the number of queues; we will reserve bold letters for vectors in $\mathbb{R}^N$, for example $\mathbf{x} = [x_n]_{1 \leq n \leq N}$. Let $|\mathbf{x}| = \max_n |x_n|$. Let $\mathbf{0}$ be the vector of all 0s and $\mathbf{1}$ the vector of all 1s. For vectors $\mathbf{u}$ and $\mathbf{v}$ let

$$\mathbf{u} \cdot \mathbf{v} = \sum_{n=1}^{N} u_n v_n, \qquad \mathbf{uv} = [u_n v_n]_{1 \leq n \leq N}$$

and let matrix multiplication take precedence over dot product so that

$$\mathbf{u} \cdot A\mathbf{v} = \sum_{n=1}^{N} u_n \left(\sum_{m=1}^{N} A_{nm} v_m \right).$$

Let A^T be the transpose of matrix A. If f is a function $\mathbb{R} \to \mathbb{R}$ then interpret $f(\mathbf{x})$ componentwise. Thus, for example,

$$\max\big(A\mathbf{w}\mathbf{q}^\alpha, 0\big) \cdot \boldsymbol{\pi} = \sum_n \max\Big(\sum_m A_{nm} w_m q_m^\alpha, 0 \Big) \pi_n.$$

Inequalities between vectors are interpreted componentwise. For $\mathcal{S} \subset \mathbb{R}^N$, let $\langle \mathcal{S} \rangle$ be the convex hull of $\mathcal{S}$. Let $1_{\{\cdot\}}$ be the indicator function, $1_{\text{true}} = 1$ and $1_{\text{false}} = 0$. Let $\mathcal{AC}$ denote the space of absolutely continuous functions $\mathbb{R}_+ \to \mathbb{R}$.

2 Model

We now define the general queueing dynamics for a multihop switched network, and the dynamics for the two scheduling policies that we analyse in this paper. We specify the dynamics in terms of fluid models; the relationship to certain stochastic queueing networks is described in Section 4.

Consider a collection of N queues, and a finite set $\mathcal{S} \subset \mathbb{R}_+^N$ of service actions, also called schedules. Assume that every queue is serviceable, i.e. for every n there exists some $\boldsymbol{\pi} \in \mathcal{S}$ such that $\pi_n > 0$.

For a multihop network let $R \in \{0, 1\}^{N \times N}$ be the routing matrix, $R_{mn} = 1$ if work served from queue m is sent to queue n, and $R_{mn} = 0$ otherwise; if $R_{mn} = 0$ for all n then work served from queue m departs the system. (Note that $R = 0$ corresponds to a single-hop network, i.e. all work leaves the system as soon as it is served.) We will assume throughout that there is no routing choice, i.e. that $\sum_n R_{mn} \in \{0, 1\}$ for each m. We will also assume throughout that routing is acyclic, i.e. that work served from some queue n never returns to queue n. This implies that the inverse $\vec{R} = (I - R^\mathsf{T})^{-1}$ exists; by considering the expansion $\vec{R} = I + R^\mathsf{T} + (R^\mathsf{T})^2 + \cdots$ it is clear that $\vec{R}_{mn} \in \{0, 1\}$ for all m and n, and that $\vec{R}_{mn} = 1$ if the

route for work injected at queue n passes through m, and 0 otherwise. For multihop networks we will additionally assume that the scheduler always has the option of not sending work downstream at every individual queue. Formally, we assume that $\mathcal{S}$ satisfies the following: if $\boldsymbol{\pi} \in \mathcal{S}$ is an allowed schedule, and $\boldsymbol{\rho} \in \mathbb{R}_+^N$ is some other vector, then

$$\text{if } \rho_n \in \{0, \pi_n\} \text{ for all } n \text{ then } \boldsymbol{\rho} \in \mathcal{S}. \tag{1}$$

Definition 1 (Queueing dynamics) *Let $\boldsymbol{\lambda} \in \mathbb{R}_+^N$. Let $\mathcal{AC}^K$ denote the space of absolutely continuous functions $\mathbb{R}_+ \to \mathbb{R}^K$, for $K \in \mathbb{N}$. Say that the triple $\mathbf{q}(\cdot) \in \mathcal{AC}^N$, $\mathbf{z}(\cdot) \in \mathcal{AC}^N$, $s(\cdot) \in \mathcal{AC}^{|\mathcal{S}|}$ is a fluid model solution for the queueing dynamics with arrival rate vector $\boldsymbol{\lambda}$ if $s(0) = 0$, $\mathbf{z}(0) = \mathbf{0}$, and the following equations are satisfied for all t:*

$$\mathbf{q}(t) = \mathbf{q}(0) + \boldsymbol{\lambda} t - (I - R^T) \sum_{\boldsymbol{\pi}} s_{\boldsymbol{\pi}}(t) \boldsymbol{\pi} + \mathbf{z}(t) \tag{2}$$

$$\sum_{\boldsymbol{\pi} \in \mathcal{S}} s_{\boldsymbol{\pi}}(t) = t \tag{3}$$

$$\text{each } s_{\boldsymbol{\pi}}(\cdot) \text{ and } z_n(\cdot) \text{ is increasing (not necessarily strictly increasing)} \tag{4}$$

$$\text{for all } n \text{ and almost all } t,\ \dot{z}_n(t) = 0 \text{ if } q_n(t) > 0 \tag{5}$$

$$\mathbf{z}(t) \leq \textstyle\sum_{\boldsymbol{\pi}} s_{\boldsymbol{\pi}}(t) \boldsymbol{\pi} \tag{6}$$

Here $\mathbf{q}(t)$ represents the vector of queue sizes at time t, $\mathbf{z}(t)$ represents the cumulative idleness up to time t, and $s_{\boldsymbol{\pi}}(t)$ represents the total amount of time spent on schedule $\boldsymbol{\pi}$ up to time t. The definition calls for these quantities to be absolutely continuous, which implies they are differentiable at almost all t. Instants at which they are all differentiable are called regular timepoints.

We next define fluid model solutions for two scheduling policies. It is most convenient to write these policies in terms of $\boldsymbol{\sigma}(t) = \sum_{\boldsymbol{\pi}} \dot{s}_{\boldsymbol{\pi}}(t) \boldsymbol{\pi}$, the vector of instantaneous service rates at time t. Whenever we write $\boldsymbol{\sigma}(t)$, we assume that t is a regular timepoint. We will refer simply to a "fluid model solution" when the context makes it clear whether we are referring to the queueing dynamics or to one or other of the two scheduling policies. With a small abuse of notation, say that $\mathbf{q}(\cdot)$ is a fluid model solution if there exist $\mathbf{z}(\cdot) \in \mathcal{AC}^N$ and $s(\cdot) \in \mathcal{AC}^{|\mathcal{S}|}$ such that $(\mathbf{q}, \mathbf{z}, s)$ is a fluid model solution.

Definition 2 (Max-weight policy) *Let $\mathbf{w} \in \mathbb{R}_{>0}^N$, and let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a differentiable and strictly increasing function for which*

$$f(0) = 0 \tag{7}$$

and for any $\mathbf{q} \in \mathbb{R}_+^N$ and $\boldsymbol{\pi} \in \mathcal{S}$, with $M(\mathbf{q}) = \max_{\boldsymbol{\rho} \in \mathcal{S}} (I - R) \mathbf{w} f(\mathbf{q}) \cdot \boldsymbol{\rho}$,

$$(I - R)\mathbf{w} f(\mathbf{q}) \cdot \boldsymbol{\pi} = M(\mathbf{q}) \quad \Longrightarrow \quad (I - R)\mathbf{w} f(\kappa \mathbf{q}) \cdot \boldsymbol{\pi} = M(\kappa \mathbf{q}) \text{ for all } \kappa \in \mathbb{R}_+. \tag{8}$$

Say that $(\mathbf{q}, \mathbf{z}, s)$ is a fluid model solution for the max-weight scheduling policy with weight function f and weights $\mathbf{w}$ if $(\mathbf{q}, \mathbf{z}, s)$ is a fluid model solution for the queueing dynamics and in addition for all regular timepoints t

$$\boldsymbol{\sigma}(t) \in \operatorname*{argmax}_{\boldsymbol{\pi} \in \langle \mathcal{S} \rangle} \mathbf{w} f(\mathbf{q}) \cdot (I - R^T) \boldsymbol{\pi}. \tag{9}$$

Note that the maximum is attained at some extreme point $\boldsymbol{\pi} \in \mathcal{S}$, since the quantity to be maximized is linear in $\boldsymbol{\pi}$. Note also that the quantity to be maximized can be rewritten as $(I - R)\mathbf{w}f(\mathbf{q}) \cdot \boldsymbol{\pi}$, i.e.

$$\sum_n \begin{cases} \big[w_n f(q_n(t)) - w_m f(q_m(t))\big]\pi_n & \text{if } m \text{ is immediately downstream of } n \\ w_n f(q_n(t))\pi_n & \text{if work served from } n \text{ leaves the system.} \end{cases}$$

The term $w_n f(q_n(t))$ may be thought of as the pressure to serve queue n, and the term $-w_m f(q_m(t))$ as the pressure not to add work to the downstream queue by serving queue n, which is why the max-weight policy is also known as backpressure.

Equation (8) says that the optimal choice of schedule is invariant when the queue sizes are rescaled. An example of a suitable weight function is $f(x) = x^\alpha$ for some $\alpha > 0$, since this is guaranteed to satisfy (8) for any $\mathcal{S}$. An example of an unsuitable weight function is $f(x) = \log(1 + x)$; it is not hard to find sets $\mathcal{S}$ such that (8) is not true with this f.

Definition 3 (α-fair policy) *Let $\alpha > 0$, and let $\mathbf{w} \in \mathbb{R}^N_{>0}$. Say that $(\mathbf{q}, \mathbf{z}, s)$ is a fluid model solution for the α-fair policy with weights $\mathbf{w}$ if $(\mathbf{q}, \mathbf{z}, s)$ is a fluid model solution for the queueing dynamics and in addition for all t*

$$\boldsymbol{\sigma}(t) \in \underset{\substack{\boldsymbol{\rho} \in \langle \mathcal{S} \rangle : \\ (I - R^T)\boldsymbol{\rho} \geq \mathbf{0}}}{\operatorname{argmax}} \ \mathbf{w}\mathbf{q}^\alpha \cdot g_\alpha\big((I - R^\mathsf{T})\boldsymbol{\rho}\big) \tag{10}$$

where

$$g_\alpha(\eta) = \begin{cases} \dfrac{\eta^{1-\alpha}}{1-\alpha} & \text{if } \alpha \neq 1 \\ \log(\eta) & \text{if } \alpha = 1 \end{cases}$$

with the added convention that $q_n^\alpha g_\alpha(0)$ is equal to 0 if $q_n = 0$ or if $\alpha < 1$, and equal to $-\infty$ if $q_n > 0$ and $\alpha \geq 1$.

It is shown in Section 6, Lemma 4, that the maximum is attained. Unlike the max-weight case however the maximum is not generally attained at an extreme point of $\langle \mathcal{S} \rangle$.

The optimization problem can be written more verbosely as follows: Given $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$ let $\boldsymbol{\rho}^{\text{up}} = R^\mathsf{T}\boldsymbol{\rho}$; then ρ_n^{up} is the sum of ρ_m over the queues m directly upstream of n. Among all $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$ such that $\rho_n \geq \rho_n^{\text{up}}$ for all n, select one to maximize

$$\sum_n w_n q_n^\alpha g_\alpha(\rho_n - \rho_n^{\text{up}}).$$

The constraint $\rho_n \geq \rho_n^{\text{up}}$ means that scheduler is not permitted to accumulate work in the middle of the network.

3 Main result

The main result of this paper is that queue sizes grow roughly like $\mathbf{q}(t) \approx t\hat{\mathbf{q}}$ where $\hat{\mathbf{q}}$ is the solution to a certain optimization problem, a different problem for each of the two scheduling policies we study. In this section we state the two optimization problems, and then the main result.

Definition 4 (Max-weight growth rates) *Consider a switched network with arrival rate* $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ *running the max-weight policy, with weight function* f *and weights* $\mathbf{w} \in \mathbb{R}_{>0}^N$. *Let* $\hat{\mathbf{q}}$ *be the unique solution to the following optimization problem, which we call* ALGP. *(Lemma 1 in Section 5 shows that* $\hat{\mathbf{q}}$ *is unique.) The optimization problem is*

$$\text{minimize } L(\mathbf{r}) = \sum_n w_n \int_0^{r_n} f(x)\,dx \quad \text{over } \mathbf{r} \in \text{FEAS}$$

where

$$\text{FEAS} = \Big\{ \mathbf{r} \in \mathbb{R}_+^N \ : \ \mathbf{r} \geq \boldsymbol{\lambda} - (I - R^\intercal)\boldsymbol{\rho} \text{ for some } \boldsymbol{\rho} \in \langle \mathcal{S} \rangle \Big\}.$$

Definition 5 (α-fair growth rates) *Consider a switched network with arrival rates* $\boldsymbol{\lambda} \in \mathbb{R}_+^N$, *running the* α*-fair policy, with* $\alpha > 0$ *and weights* $\mathbf{w} \in \mathbb{R}_{>0}^N$. *Let* $\hat{\mathbf{q}}$ *be the unique solution to the following optimization problem, which we call* DEP. *(Lemma 4 in Section 6 shows that it is unique.) The optimization problem is*

$$\text{minimize } H(\mathbf{q}) = \sum_n w_n \int_0^{q_n} \Big(\frac{\lambda_n}{x} - 1\Big)^{-\alpha} dx \quad \text{over } \mathbf{q} \in \text{FEAS}$$

with the convention that the objective is ∞ *if there is some* n *with* $q_n > \lambda_n$, *and terms with* $\lambda_n = q_n = 0$ *contribute* 0 *to the sum. The feasible set is*[2]

$$\text{FEAS} = \Big\{ \mathbf{r} \in \mathbb{R}_+^N \ : \ \mathbf{r} \geq \boldsymbol{\lambda} - (I - R^\intercal)\boldsymbol{\rho} \text{ for some } \boldsymbol{\rho} \in \langle \mathcal{S} \rangle, \\ \text{and } r_n = 0 \text{ for all queues } n \text{ where } \lambda_n = 0 \Big\}.$$

We now state the main result. There are two versions, one for max-weight and one for α-fair.

Theorem 1 (Max-weight version) *Consider a switched network running the max-weight policy. Let* $\mathbf{q}(t)$ *be any fluid model solution, and let* $\hat{\mathbf{q}}$ *be as in Definition 4. Then* $\mathbf{q}(t)/t \to \hat{\mathbf{q}}$ *as* $t \to \infty$. *Furthermore, for any* $c > 0$ *the convergence is uniform over the set of fluid model solutions for which* $|\mathbf{q}(0)| \leq c$. *Furthermore, if* $\mathbf{q}(0) = \mathbf{0}$ *then* $\mathbf{q}(t) = t\hat{\mathbf{q}}$.

Theorem 2 (α-fair version) *Consider a switched network running the* α*-fair policy. Let* $\mathbf{q}(t)$ *be any fluid model solution satisfying*

$$q_n(0) = 0 \quad \text{if} \quad \lambda_n = 0, \tag{11}$$

and let $\hat{\mathbf{q}}$ *be as in Definition 5. Then* $\mathbf{q}(t)/t \to \hat{\mathbf{q}}$ *as* $t \to \infty$. *Furthermore, for any* $c > 0$ *the convergence is uniform over the set of fluid model solutions satisfying* (11) *for which* $|\mathbf{q}(0)| \leq c$. *Furthermore, if* $\mathbf{q}(0) = \mathbf{0}$ *then* $\mathbf{q}(t) = t\hat{\mathbf{q}}$.

[2] We have deliberately reused the name FEAS, even though it is defined differently for maxweight and for α-fair. We will also reuse the name ALGP for an optimization problem used in the proofs for the α-fair policy, similar to ALGP in Definition 4. We are reusing names like this since there are substantial parts of the proofs which are nearly identical for maxweight and for α-fair, differing only in which version of FEAS and ALGP they refer to.

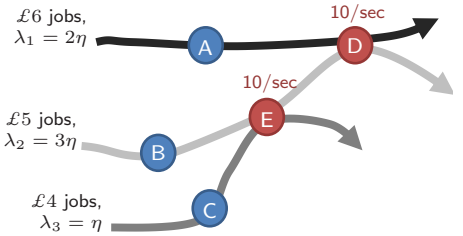


Fig. 1 Nodes A, B and C have lots of spare capacity, but nodes D and E can only serve at rate 10 jobs per second. The arrival rates are $\boldsymbol{\lambda} = (2\eta, 3\eta, \eta)$. When the system becomes overloaded, at $\eta \geq 2$, the revenue-maximizing action is to discard some of the jobs worth £5 in favour of jobs worth £4 at node E.

4 Applications

In this section we give several applications of our main result. We start in Section 4.1 with a stylized toy model of a web data center, to illustrate congestion collapse and to show how revenue can be maximized by using max-weight scheduling with appropriate weights. In Section 4.2 we summarize the model of an input-queued switch introduced by Dai and Prabhakar [5], and in Section 4.3 give explicit calculations for congestion collapse in a 2×2 switch running max-weight. In Section 4.4 we generalize these two examples, and show that max-weight gets arbitrarily close to optimal throughput for any switched network, by chosing an appropriate weight function. Finally in Section 4.5 we describe the model for α-fair sharing of bandwidth in the Internet.

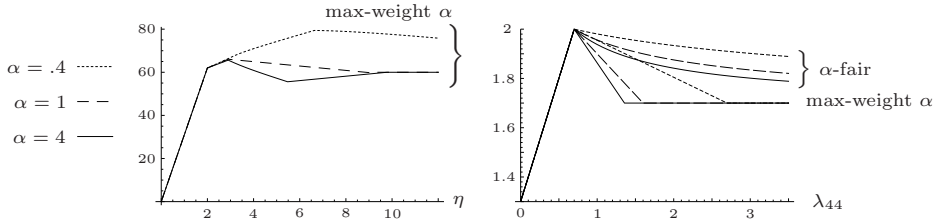
4.1 Congestion collapse in a data center

A data center consists of a network of machines, some of them running web server software, some running database software. A page request from an outside user is first directed to a web server machine, which may then trigger internal requests to several database machines, each of which may itself trigger further internal requests. In order to provide a complete response to the page request, all of these internal requests need to be served.

Figure 1 depicts a toy example of a data center which handles three types of page request, each of which needs to be processed at a web server (A, B or C) and also at database machines (D and E). It is easy to see how congestion collapse might occur: if both database machines are overloaded, then machine E could expend service on £5 jobs only for them to be dropped when they reach machine D. That service would have been better spent on £4 jobs. A real example of a distributed database exhibiting congestion collapse is given in [12].

In a data center, some requests may be more valuable than others, and so maximizing revenue may be a more natural goal than maximizing throughput. Suppose in this example that the data center earns £6 when it completely serves a request labelled £6, i.e. when the request has been served by both A and D, and similarly for the other request classes. Instead of congestion collapse, the concern is now revenue collapse. What scheduling algorithms at D and E can avoid this?

To be concrete, suppose A, B and C forward requests immediately, and let there be four queues, one for each request class at each database machine, call them q_{D5}, q_{D6}, q_{E4} and



(a) data center (b) input-queued switch

Fig. 2 Two examples of congestion collapse. Figure (a) shows revenue as a function of load, for the data center example in Section 4.1, when running the max-weight algorithm with weight function $f(q) = q^\alpha$. Figure (b) shows throughput as a function of load, for the 2×2 input-queued switch in Section 4.3, when running either α-fair scheduling or max-weight scheduling with weight function $f(q) = q^\alpha$.

q_{E5}. Revenue earned by time t is equal to the total potential revenue brought by arriving requests, minus revenue not earned because requests have not yet left the system. Therefore, in order to maximize revenue, we want to

$$\text{minimize} \quad 6\,q_{D6} + 5\,(q_{E5} + q_{D5}) + 4\,q_{E4}. \tag{12}$$

Suppose we run max-weight with weight function $f(q) = q^\alpha$, $\alpha > 0$, and let the weight of each queue be the revenue associated with requests in that queue. Under this policy, machine E, for example, when it is ready to serve a new request, will serve q_{E5} if $5q_{E5}^\alpha - 5q_{D5}^\alpha > 4q_{E4}^\alpha$, it will serve q_{E4} if the inequality is reversed, and its choice is arbitrary if the two are equal. Theorem 1 says that when the system is overloaded, the queue sizes grow like $\mathbf{q}(t) = t\hat{\mathbf{q}}$, where $\hat{\mathbf{q}}$ solves the optimization problem

$$\begin{aligned} &\text{minimize} \quad \frac{1}{1+\alpha}\Big(6\,r_{D6}^{1+\alpha} + 5\,(r_{E5}^{1+\alpha} + r_{D5}^{1+\alpha}) + 4\,r_{E4}^{1+\alpha}\Big) \\ &\text{over all feasible growth rates } \mathbf{r}. \end{aligned}$$

Figure 2(a) shows revenue rate as a function of η, for three different values of α, obtained by numerically solving this optimization. The highest possible revenue rate is £100, achievable for $\eta \geq 10$.

If we choose $\alpha \approx 0$, then the objective function that defines $\hat{\mathbf{q}}$ is very close to the equation for lost revenue, equation (12), so the max-weight policy is very close to revenue-maximizing. (This is made precise in Section 4.4.) In effect, information about downstream congestion is propagated upstream by means of queue sizes, telling upstream nodes to hold back work which won't be able to be served downstream. The max-weight policy does not need to estimate arrival rates in order to achieve this: its actions are based purely on queue size.

4.2 Fluid model for an input-queued switch

Dai and Prabhakar [5] introduced a fluid model for input-queued switches running max-weight. A switch is the core of an Internet router, and the input-queued architecture is commercially popular. We will now describe the stochastic model, and their fluid model,

and explain how it relates to ours, and how to interpret our main result in the original stochastic system.

Consider a collection of $N = M^2$ queues operating in slotted time; in this application it is most natural to consider the queue lengths to be a matrix in $\mathbb{R}_+^{M\times M}$ rather than a vector in $\mathbb{R}_+^N$. At the beginning of each timeslot, a (random) integer number of packets arrive, and there may be arrivals to any queue. Then a service action $\boldsymbol{\pi}$ is chosen from the set $\mathcal{S} \subset \{0,1\}^{M\times M}$ consisting of all $M!$ permutation matrices. During the timeslot, $\boldsymbol{\pi}$ is the offered service to each queue, and served work leaves the system at the end of the timeslot. The natural questions are what scheduling policy should be used, and what the resulting performance of the system is. Dai and Prabhakar [5] analyse the stability of two different policies, one of which is the max-weight policy with weights $\mathbf{w} = \mathbf{1}$ and weight function $f(q) = q$, which is our focus here.

Let $\mathbf{Q}(t)$ be the matrix of queue sizes at timeslot $t \in \mathbb{N}$, let $S_{\boldsymbol{\pi}}(t)$ be the total number of timeslots up to and including t in which action $\boldsymbol{\pi}$ has been chosen, and let $\mathbf{D}(t)$ be the number of departures from each queue by the end of timeslot t. Extend these to be continuous functions of $t \in \mathbb{R}_+$ by polygonalization, and define fluid-scaled versions $\mathbf{q}^r(t) = \mathbf{Q}(rt)/r$ etc. Any weak limit as $r \to \infty$ of these processes is called a fluid limit. It is shown in [5] that every fluid limit satisfies the following fluid model:

$$\begin{aligned}
&\mathbf{q}(t) = \mathbf{q}(0) + \boldsymbol{\lambda} t - \mathbf{d}(t) \geq \mathbf{0}\\
&\dot{d}_{ij}(t) = \sum_{\boldsymbol{\pi}\in\mathcal{S}} \pi_{ij} \dot{s}_{\boldsymbol{\pi}}(t), \text{ if } q_{ij}(t) > 0\\
&s_{\boldsymbol{\pi}}(\cdot) \text{ is non-decreasing, and } \sum_{\boldsymbol{\pi}\in\mathcal{S}} s_{\boldsymbol{\pi}}(t) = t\\
&\dot{s}_{\boldsymbol{\pi}}(t) = 0 \text{ if } \boldsymbol{\pi}\cdot\mathbf{q}(t) < \max_{\boldsymbol{\rho}\in\mathcal{S}} \boldsymbol{\rho}\cdot\mathbf{q}(t).
\end{aligned}$$

By rewriting these equations in terms of the cumulative idleness process $\mathbf{z}(t)$ rather than the cumulative departure process $\mathbf{d}(t) = \sum_{\boldsymbol{\pi}} s_{\boldsymbol{\pi}}(t)\boldsymbol{\pi} - \mathbf{z}(t)$, and adding the natural constraint $\dot{\mathbf{d}}(t) \geq \mathbf{0}$, we obtain our fluid model equations (2)–(6) and (9), with routing matrix $R = 0$. The general fluid equations, including multihop and arbitrary weight functions, but restricted to weights $\mathbf{w} = \mathbf{1}$, are described by e.g. [15].

Theorem 1 says that, if $\mathbf{q}(0) = \mathbf{0}$, then the unique solution to the fluid model equations is $\mathbf{q}(t) = t\hat{\mathbf{q}}$. Therefore every fluid limit point of the sequence $\mathbf{Q}(r)/r$ is equal to $\hat{\mathbf{q}}$, hence $\mathbf{Q}(r)/r \to \hat{\mathbf{q}}$ almost surely.

To our knowledge, no one has studied input-queued switches running the α-fair algorithm. We conjecture that the α-fair fluid dynamics are obtained as the fluid limit of the following algorithm: (1) each timeslot t, compute $\boldsymbol{\sigma}(t) \in \langle\mathcal{S}\rangle$ to solve (10), then (2) write $\boldsymbol{\sigma}(t)$ as a convex combination of elements of $\mathcal{S}$, $\boldsymbol{\sigma}(t) = \sum_{\boldsymbol{\pi}} x_{\boldsymbol{\pi}}\boldsymbol{\pi}$, then (3) pick some service action at random, with the probability of choosing $\boldsymbol{\pi}$ equal to $x_{\boldsymbol{\pi}}$. The service actions at each timeslot must be chosen independently, conditional on the queue sizes.

4.3 Congestion collapse in a 2×2 input-queued switch

Here is a worked example demonstrating congestion collapse in an input-queued switch. We have already seen congestion collapse in the multihop data center example, but this

example shows that congestion collapse can also happen in a single-hop network. In essence, the scheduler is 'tricked' into choosing a bad combination of queues to serve.

Let the queue sizes of the four queues in a 2×2 input-queued switch be q_{11}, q_{12}, q_{21}, q_{22}. The two possible service actions are "serve queues (11) and (22)" and "serve queues (12) and (21)". The max-weight policy is to serve queues (11) and (22) if $q_{11} + q_{22} > q_{12} + q_{21}$, to serve queues (12) and (21) if $q_{11} + q_{22} < q_{12} + q_{21}$, and to choose randomly if there is equality. Let the arrival rate vector be

$$\boldsymbol{\lambda} = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & \lambda_{22} \end{pmatrix}.$$

Theorem 1 says that queue sizes grow linearly, $\mathbf{q}(t) = t\hat{\mathbf{q}}$ where $\hat{\mathbf{q}}$ is specified in Definition 4. Therefore the cumulative departure process is $\mathbf{d}(t) = \boldsymbol{\lambda} t - \mathbf{q}(t) = (\boldsymbol{\lambda} - \hat{\mathbf{q}})t$ and so the throughput, i.e. the instantaneous net departure rate, is

$$\text{throughput} = \dot{\mathbf{d}}(t) \cdot \mathbf{1} = (\boldsymbol{\lambda} - \hat{\mathbf{q}}) \cdot \mathbf{1}.$$

After calculating $\hat{\mathbf{q}}$ as specified in Definition 4, we find

$$\text{throughput} = \begin{cases} 1.3 + \lambda_{22} & \text{if } \lambda_{22} \le 0.7 \\ 2 - (\lambda_{22} - 0.7)/3 & \text{if } 0.7 \le \lambda_{22} \le 1.6 \\ 1.7 & \text{if } \lambda_{22} \ge 1.6. \end{cases}$$

This shows that an increase in offered load beyond $\lambda_{22} > 0.7$ can lead to a loss in throughput. Suppose instead we use the generalized max-weight policy with weight function $f(q) = q^\alpha$, $\alpha > 0$. In other words, we use the policy "serve queues (11) and (22) if $q_{11}^\alpha + q_{22}^\alpha > q_{12}^\alpha + q_{21}^\alpha$, and serve queues (12) and (21) if $q_{11}^\alpha + q_{22}^\alpha < q_{12}^\alpha + q_{21}^\alpha$". Now the throughput is

$$\text{throughput} = \begin{cases} 1.3 + \lambda_{22} & \text{if } \lambda_{22} \le 0.7 \\ 2 - (\lambda_{22} - 0.7)/(1 + 2^{1/\alpha}) & \text{if } 0.7 \le \lambda_{22} \le 1 + 0.3 \times 2^{1/\alpha} \\ 1.7 & \text{if } \lambda_{22} \ge 1 + 0.3 \times 2^{1/\alpha}. \end{cases}$$

This shows that, for any $\lambda_{22} > 0.7$, the throughput increases to its maximum possible value, namely 2, as α decreases to 0.

The α-fair policy described in the previous section is to choose the action "serve queues (11) and (22)" with probability

$$\frac{(q_{11}^\alpha + q_{22}^\alpha)^{1/\alpha}}{(q_{11}^\alpha + q_{22}^\alpha)^{1/\alpha} + (q_{12}^\alpha + q_{21}^\alpha)^{1/\alpha}}.$$

The throughput for this algorithm can be calculated explicitly, with some work. As with maxweight, throughput increases up to $\lambda_{22} = 0.7$ and thereafter it decreases, converging to 1.7 as $\lambda_{22} \to \infty$. Throughputs for both algorithms, for $\alpha \in \{.4, 1, 4\}$, are shown in Figure 2(b).

4.4 Near-optimality of max-weight policy

In both examples, the data center in Section 4.1 and the input-queued switch in Section 4.3, we saw that the max-weight policy with weight function $f(q) = q^\alpha$ has useful properties as $\alpha \to 0$. We now make a general statement about near-optimality of max-weight for any switched network.

Theorem 3 *Consider a switched network with N queues, and let $\mathbf{v} \in \mathbb{R}_+^N$ be some vector of weights. Let the queues start empty. Pick $\alpha > 0$, and let $\mathbf{q}(t)$ be the fluid model solution for the max-weight algorithm with weight function $f(q) = q^\alpha$ and weight vector $\mathbf{w} = \mathbf{v}^{1+\alpha}$. Also, let $\mathbf{r}(t)$ be a fluid model solution for the queueing dynamics, for any scheduling algorithm. Then*

$$\mathbf{v}\cdot\mathbf{q}(t) \le N^{\alpha/(1+\alpha)}\mathbf{v}\cdot\mathbf{r}(t).$$

The constant factor $N^{\alpha/(1+\alpha)}$ can be made arbitrarily close to 1 by choosing α sufficiently small. In Section 4.1, maximum revenue is achieved by a scheduling policy that minimizes $\mathbf{v}\cdot\mathbf{r}(t)$, where v_n is the revenue earned when a job in queue n eventually leaves the network. In Section 4.3, maximum throughput is achieved by a scheduling policy that minimizes $\mathbf{1}\cdot\mathbf{r}(t)$. Thus, in both these examples, max-weight scheduling (with the above choice of weights, and with α sufficiently small) is near-optimal.

Proof First we argue that $\mathbf{r}(t)/t \in$ FEAS, where FEAS is given in Definition 4. From (2), and using the fact that $\mathbf{z}(\cdot)$ starts at $\mathbf{0}$ and is increasing by (4), $\mathbf{r}(t) \ge \boldsymbol{\lambda}t - (I - R^\mathsf{T})\sum_{\boldsymbol{\pi}} \boldsymbol{\pi} s_{\boldsymbol{\pi}}(t)$. Dividing by t, $\mathbf{r}(t)/t \ge \boldsymbol{\lambda}t - (I - R^\mathsf{T})\boldsymbol{\rho}$ where $\boldsymbol{\rho} = \sum_{\boldsymbol{\pi}} \boldsymbol{\pi} s_{\boldsymbol{\pi}}(t)/t$, and $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$ by (3). Hence $\mathbf{r}(t)/t \in$ FEAS.

Now, by Theorem 1, $\mathbf{q}(t) = t\hat{\mathbf{q}}$ where $\hat{\mathbf{q}}$ solves the optimization problem

$$\text{minimize } \mathbf{w}\cdot\mathbf{r}^{1+\alpha} \quad \text{over } \mathbf{r} \in \text{FEAS}\,.$$

Since $\mathbf{r}(t)/t \in$ FEAS, we deduce

$$\mathbf{w}\cdot\big(\mathbf{q}(t)/t\big)^{1+\alpha} \le \mathbf{w}\cdot\big(\mathbf{r}(t)/t\big)^{1+\alpha}. \tag{13}$$

It is a standard result about norms, from Hölder's inequality and Jensen's inequality, that for any $\mathbf{x} \in \mathbb{R}_+^N$ and $\beta > 1$,

$$\frac{1}{N^{1-1/\beta}}\,\mathbf{x}\cdot\mathbf{1} \le (\mathbf{x}^\beta\cdot\mathbf{1})^{1/\beta} \le \mathbf{x}\cdot\mathbf{1}.$$

Multipling each side of (13) by $t^{1+\alpha}$ and applying the standard result about norms with $\beta = 1+\alpha$, first with $x_n = w_n^{1/(1+\alpha)} q_n(t)$ then with $x_n = w_n^{1/(1+\alpha)} r_n(t)$,

$$\begin{aligned}\frac{1}{N^{\alpha/(1+\alpha)}}\mathbf{w}^{1/(1+\alpha)}\cdot\mathbf{q}(t) &\le \big(\mathbf{w}\cdot\mathbf{q}(t)^{1+\alpha}\big)^{1/(1+\alpha)} \\ &\le \big(\mathbf{w}\cdot\mathbf{r}(t)^{1+\alpha}\big)^{1/(1+\alpha)} \le \mathbf{w}^{1/(1+\alpha)}\cdot\mathbf{r}(t).\end{aligned}$$

Rewriting in terms of $\mathbf{v} = \mathbf{w}^{1/(1+\alpha)}$ we obtain the result. □

4.5 Fluid model and congestion collapse for bandwidth sharing

Roberts and Massoulie [14] introduced a model for bandwidth-sharing in the Internet. They took there to be a finite set J of links, and for each link j an associated capacity $C_j \geq 0$, and a finite set $R = \{r_1, \ldots, r_N\}$ of routes where each route r_n is a subset of J. At every instant in time t, there is a certain number $x_n(t)$ of active flows on link n. These flows receive service at a certain rate, which depends only on the number of active flows: let $\sigma_n^*(\mathbf{x})/x_n$ be the service rate for each flow on route n when $\mathbf{x}$ gives the number of active flows on each link. We can think of the Internet's congestion control policy (TCP) as selecting a particular service rate vector $\boldsymbol{\sigma}^*(\mathbf{x})$ that satisfies the capacity constraint $A\boldsymbol{\sigma}^*(\mathbf{x}) \leq C$ where $A_{jn} = 1_{j \in r_n}$. Let $\mathcal{S}$ be the set of extreme points of $\{\boldsymbol{\rho} \in \mathbb{R}_+^N : A\boldsymbol{\rho} \leq C\}$, of which there are finitely many; then the constraint can be written $\boldsymbol{\sigma}^*(\mathbf{x}) \in \langle \mathcal{S} \rangle$.

Kelly and Williams [11] introduced fluid model equations for this system for the case that flow sizes have an exponential distribution, say with mean $1/\mu_n$ on route n, new flows arrive on route r_n as a Poisson process $\lambda_n \mu_n$, and where the service rate vector is chosen to be α-fair. If we rewrite their equations in terms of $q_n(t) = x_n(t)/\mu_n$, we obtain (2)–(6) and (10). The bandwidth-sharing model corresponds to a single-hop switched network, i.e. $R = 0$. The fluid model has been generalized to allow for general distributions for flow size [8], but this is a level of generality which we do not address in this paper.

Egorova et al [6] have studied overload in the bandwidth-sharing model, under the α-fair policy. They allow general flow size distributions. They formulate an optimization problem, 'Problem Q', upon which we based our optimization problem DEP in Definition 5. They prove that this optimization problem has a unique solution $\hat{\mathbf{q}}$, and that $\mathbf{q}(t) = \hat{\mathbf{q}}t$ is a feasible solution to the fluid model dynamics, and that it is the unique fluid model solution with linear trajectories (but they do not prove that it is the unique fluid model solution whereas we do).

We have investigated various examples numerically and seen congestion collapse, as well as convergence to optimal throughput as $\alpha \to 0$. However, we do not have any general results along the lines of Theorem 3.

5 Proofs for the max-weight policy

The proof consists of Lemma 1 showing that $\hat{\mathbf{q}}$ is unique, Lemma 2 showing that $\mathbf{q}(t)/t \in$ FEAS for any fluid model solution $\mathbf{q}(\cdot)$, Lemma 3 showing that the function $L(\cdot)$ appearing in ALGP is a Lyapunov function, and finally a proof of the main theorem.

Lemma 1 *For the network specified in Definition 4,* ALGP *has a unique solution.*

Proof The set FEAS is certainly non-empty; pick any $\mathbf{r}^* \in$ FEAS. The solution to ALGP must be at least as good as $\mathbf{r}^*$, so we may as well restrict the optimization to be over $D = \{\mathbf{r} \in \text{FEAS} : L(\mathbf{r}) \leq L(\mathbf{r}^*)\}$. Since f is ≥ 0 and strictly increasing, $L(\mathbf{r}^*) \leq N|\mathbf{r}^*| f\big(|\mathbf{r}^*|\big) < \infty$ and $L(\mathbf{r}) \to \infty$ as $|\mathbf{r}| \to \infty$; therefore D is bounded. It is easy to check that D is also closed and convex. The objective function is strictly convex, since f is ≥ 0 and strictly increasing and $\mathbf{w} > \mathbf{0}$ componentwise. Since the optimization is of a strictly convex function over a closed, convex, bounded domain, there is a unique optimum. □

Lemma 2 *For any fluid model solution of the queue dynamics with arrival rate* $\boldsymbol{\lambda}$, $\mathbf{q}(t)/t \in$ FEAS *for all* $t > 0$.

Proof Let $\mathbf{q}(\cdot)$ be any fluid model solution. From (2), and using the fact that $\mathbf{z}(\cdot)$ is increasing by (4) and starts at $\mathbf{0}$, $\mathbf{q}(t) \geq \boldsymbol{\lambda} t - (I - R^{\mathsf{T}}) \sum_{\boldsymbol{\pi}} \boldsymbol{\pi} s_{\boldsymbol{\pi}}(t)$. Dividing by t, $\mathbf{q}(t)/t \geq \boldsymbol{\lambda} t - (I - R^{\mathsf{T}})\boldsymbol{\rho}$ where $\boldsymbol{\rho} = \sum_{\boldsymbol{\pi}} \boldsymbol{\pi} s_{\boldsymbol{\pi}}(t)/t$, and $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$ by (3). □

Lemma 3 *For the network specified in Definition 4,*

$$t \frac{d}{dt} L\Big(\frac{\mathbf{q}(t)}{t}\Big) \leq L(\hat{\mathbf{q}}) - L\Big(\frac{\mathbf{q}(t)}{t}\Big) \leq 0.$$

Proof For the first inequality, the drift we want to bound is

$$t \frac{d}{dt} L\Big(\frac{\mathbf{q}(t)}{t}\Big) = \mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot \Big(\frac{d\mathbf{q}(t)}{dt} - \frac{\mathbf{q}(t)}{t}\Big).$$

For each queue n, either $q_n(t) > 0$ in which case $\dot{q}_n(t) = \lambda_n - [(I - R^{\mathsf{T}})\boldsymbol{\sigma}(t)]_n$ by (2) and (5), or $q_n(t) = 0$ in which case $f(q_n(t)/t) = 0$ by (7), hence

$$t \frac{d}{dt} L\Big(\frac{\mathbf{q}(t)}{t}\Big) = \mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot \Big(\boldsymbol{\lambda} - (I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) - \frac{\mathbf{q}(t)}{t}\Big). \tag{14}$$

By the max-weight property (9) and the scale-invariance property (8),

$$\mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot (I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) = \max_{\boldsymbol{\pi} \in \mathcal{S}} \mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot (I - R^{\mathsf{T}})\boldsymbol{\pi}.$$

Now, the optimum in ALGP is attained at $\hat{\mathbf{q}} \in$ FEAS, hence for some $\hat{\boldsymbol{\rho}} \in \langle \mathcal{S} \rangle$

$$\hat{\mathbf{q}} \geq \boldsymbol{\lambda} - (I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}. \tag{15}$$

This $\hat{\boldsymbol{\rho}}$ is a feasible choice for the scheduling policy, while $\boldsymbol{\sigma}(t)$ is the optimal choice, so

$$\mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot (I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) \geq \mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot (I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}.$$

Substituting this into (14), and using (15),

$$t \frac{d}{dt} L\Big(\frac{\mathbf{q}(t)}{t}\Big) \leq \mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot \Big(\boldsymbol{\lambda} - (I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}} - \frac{\mathbf{q}(t)}{t}\Big) \leq \mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot \Big(\hat{\mathbf{q}} - \frac{\mathbf{q}(t)}{t}\Big).$$

Finally, by convexity of $L(\cdot)$,

$$\begin{aligned} L(\hat{\mathbf{q}}) &\geq L\Big(\frac{\mathbf{q}(t)}{t}\Big) + \nabla L\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot \Big(\hat{\mathbf{q}} - \frac{\mathbf{q}(t)}{t}\Big) \\ &= L\Big(\frac{\mathbf{q}(t)}{t}\Big) + \mathbf{w} f\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot \Big(\hat{\mathbf{q}} - \frac{\mathbf{q}(t)}{t}\Big) \end{aligned}$$

and rearranging this gives the desired result.

The second inequality in the statement of the theorem is straightforward, from the observation that $\mathbf{q}(t)/t \in$ FEAS by Lemma 2 and $\hat{\mathbf{q}}$ is optimal for ALGP. □

We can now prove the main result.

Proof of Theorem 1, first two claims. The first claim of the theorem, that $\mathbf{q}(t)/t \to \hat{\mathbf{q}}$, follows trivially from the second claim, that convergence is uniform: simply set $c = |\mathbf{q}(0)|$. For the second claim, pick $c > 0$ and $\varepsilon > 0$, and define

$$\begin{aligned}
\mathbf{q}^{\max} &= \mathbf{1}\big(c + |\boldsymbol{\lambda}| + \max_{\boldsymbol{\pi}\in\mathcal{S}} |R^{\mathsf{T}}\boldsymbol{\pi}|\big) \\
\mathcal{D}_c &= \Big\{\mathbf{r} \in \mathbb{R}_+^N \,:\, L(\mathbf{r}) \le L(\mathbf{q}^{\max}) \text{ and } \mathbf{r} \in \text{FEAS}\Big\} \\
\mathcal{I}_\varepsilon &= \Big\{\mathbf{r} \in \mathbb{R}_+^N \,:\, |\mathbf{r} - \hat{\mathbf{q}}| < \varepsilon\Big\} \\
K_{\varepsilon,c} &= \inf\big\{L(\mathbf{r}) - L(\hat{\mathbf{q}}) \,:\, \mathbf{r} \in \mathcal{D}_c \setminus \mathcal{I}_\varepsilon\big\} \\
\mathcal{K}_{\varepsilon,c} &= \Big\{\mathbf{r} \in \mathbb{R}_+^N \,:\, L(\mathbf{r}) - L(\hat{\mathbf{q}}) < K_{\varepsilon,c}\Big\} \\
H_{\varepsilon,c} &= \exp\big(L(\mathbf{q}^{\max})/K_{\varepsilon,c}\big).
\end{aligned}$$

Let $\mathbf{q}(\cdot)$ be any fluid model solution with $|\mathbf{q}(0)| < c$. We will show in a moment that $\mathbf{q}(t)/t$ enters $\mathcal{K}_{\varepsilon,c}$ within time $H_{\varepsilon,c}$. By Lemma 3, $L(\mathbf{q}(t)/t)$ is non-increasing, therefore $\mathbf{q}(t)/t \in \mathcal{K}_{\varepsilon,c}$ for all times $t \ge H_{\varepsilon,c}$. By construction of $\mathcal{K}_{\varepsilon,c}$, if $\mathbf{r} \in \mathcal{K}_{\varepsilon,c}$ then either $\mathbf{r} \in \mathcal{I}_\varepsilon$ or $\mathbf{r} \notin \mathcal{D}_c$ (or both). We will show in a moment that $\mathbf{q}(t)/t \in \mathcal{D}_c$ for all $t \ge 1$. Thus $\mathbf{q}(t)/t \in \mathcal{I}_\varepsilon$ for all $t \ge \max(H_{\varepsilon,c}, 1)$. Since $H_{\varepsilon,c}$ does not depend on $\mathbf{q}(\cdot)$, we have established uniform convergence. We have two things left to show:

Proof that $\mathbf{q}(t)/t \in \mathcal{D}_c$ for $t \ge 1$. From (2) and (6), $\mathbf{q}(1) \le \mathbf{q}(0) + \boldsymbol{\lambda} + R^{\mathsf{T}} \sum_{\boldsymbol{\pi}} s_{\boldsymbol{\pi}}(1)\boldsymbol{\pi}$ and this is $\le \mathbf{q}^{\max}$. Since L is increasing componentwise, $L(\mathbf{q}(1)/1) = L(\mathbf{q}(1)) \le L(\mathbf{q}^{\max})$. By Lemma 3, $L(\mathbf{q}(t)/t)$ is non-increasing hence $L(\mathbf{q}(t)/t) \le L(\mathbf{q}^{\max})$ for all $t \ge 1$. Finally, Lemma 2 shows that $\mathbf{q}(t)/t \in \text{FEAS}$ for all $t > 0$. Hence the claimed result.

Proof that $\mathbf{q}(t)/t$ enters $\mathcal{K}_{\varepsilon,c}$ within time $H_{\varepsilon,c}$. Observe that (i) $L(\mathbf{q}^{\max})$ is finite; (ii) $\mathcal{D}_c$ is bounded because L is convex and increasing in each component, and it is closed because L is continuous and the feasibility constraint in FEAS is continuous, hence $\mathcal{D}_c$ is compact; (iii) the infimum in $K_{\varepsilon,c}$ is of a continuous function over a compact set, hence it is attained say at some $\hat{\mathbf{r}}$, (iv) $L(\hat{\mathbf{r}}) > L(\hat{\mathbf{q}})$ because $\hat{\mathbf{r}}$ is feasible for ALGP and $\hat{\mathbf{q}}$ is the unique optimum, hence $K_{\varepsilon,c} > 0$.

Now, consider $\ell(u) = L(\mathbf{q}(e^u)/e^u)$ for $u \ge 0$. Using Lemma 3 and the chain rule,

$$\frac{d}{du}\ell(u) \le -\Big\{L\Big(\frac{\mathbf{q}(e^u)}{e^u}\Big) - L(\hat{\mathbf{q}})\Big\} \le 0.$$

So $\ell(\cdot)$ is non-increasing, and $d\ell(u)/du \le -K_{\varepsilon,c}$ for all the time that $\mathbf{q}(e^u)/e^u \notin \mathcal{K}_{\varepsilon,c}$. We know from the previous claim that $\ell(0) = L(\mathbf{q}(1)) \le L(\mathbf{q}^{\max})$, and it is clear by definition that $g(u) \ge 0$ for all u. Hence $\mathbf{q}(e^u)/e^u$ must enter $\mathcal{K}_{\varepsilon,c}$ within the interval $u \in [0, L(\mathbf{q}^{\max})/K_{\varepsilon,c}]$, hence the claimed result.

Proof of Theorem 1, third claim. By Theorem 1, given $\varepsilon > 0$ there is some time H_ε such that $|\mathbf{q}(t)/t - \hat{\mathbf{q}}| < \varepsilon$ for all $t \ge H_\varepsilon$ and all fluid model solutions $\mathbf{q}(\cdot)$ with $\mathbf{q}(0) = \mathbf{0}$.

Now let $\mathbf{q}(\cdot)$ be any such fluid model solution, and suppose $\mathbf{q}(t) \ne t\hat{\mathbf{q}}$ for some $t > 0$, say $|\mathbf{q}(t_0)/t_0 - \hat{\mathbf{q}}| = \varepsilon > 0$. Consider the rescaled sample path $\tilde{\mathbf{q}}(t) = \mathbf{q}(t\kappa)/\kappa$ where $\kappa = t_0/H_\varepsilon$: this is chosen so that

$$\Big|\frac{\tilde{\mathbf{q}}(H_\varepsilon)}{H_\varepsilon} - \hat{\mathbf{q}}\Big| = \varepsilon. \tag{16}$$

Let $\mathbf{z}(\cdot)$ and $s(\cdot)$ be idleness and service processes associated with the fluid model solution $\mathbf{q}(\cdot)$, and consider rescaled versions of these: $\tilde{\mathbf{z}}(t) = \mathbf{z}(\kappa t)/\kappa$ and $\tilde{s}(t) = s(\kappa t)/t$. These all satisfy the fluid model equations, hence $\tilde{\mathbf{q}}$ is a fluid model solution. Furthermore $\tilde{\mathbf{q}}(0) = \mathbf{q}(0) = \mathbf{0}$, so by choice of H_ε it must be that $|\tilde{\mathbf{q}}(t)/t - \hat{\mathbf{q}}| < \varepsilon$ for all $t \geq H_\varepsilon$. This contradicts (16), hence the supposition that $\mathbf{q}(t_0)/t_0 \neq \hat{\mathbf{q}}$ is false. □

6 Proofs for the α-fair policy

We begin with a basic lemma which shows that the two optimization problems we have defined for α-fair scheduling, namely the optimization problem which specifies $\boldsymbol{\sigma}(t)$ and the optimization problem which specifies $\hat{\mathbf{q}}$, make sense.

Lemma 4 *i. The problem* DEP *in Definition 5 has a unique solution, call it* $\hat{\mathbf{q}}$*. Furthermore,* $\hat{\mathbf{q}} \leq \boldsymbol{\lambda}$*, and* $\hat{q}_n < \lambda_n$ *if* $\lambda_n > 0$*.*
ii. In the definition of the α*-fair policy, the maximum in* (10) *is attained.*

Proof of (i). We will first argue that there exists $\boldsymbol{\rho}^* \in \langle \mathcal{S} \rangle$ such that $(I - R^\mathsf{T})\boldsymbol{\rho}^* > \mathbf{0}$ componentwise. We have two arguments, one for single-hop networks and one for multihop.

First the case of a single-hop network. We assumed that every queue is serviceable, so for each queue n we can pick $\boldsymbol{\pi}^n \in \mathcal{S}$ such that $\pi^n_n > 0$. Let $\boldsymbol{\rho}^* = N^{-1}\sum_n \boldsymbol{\pi}^n$; then $\boldsymbol{\rho}^* = (I - R^\mathsf{T})\boldsymbol{\rho}^* > \mathbf{0}$ componentwise since each $\boldsymbol{\pi}^n$ is ≥ 0 and $R = 0$.

In the case of a multihop network, define $\boldsymbol{\pi}^n$ instead by $\pi^n_k = \varepsilon 1_{n=k}$. We assumed that every queue is serviceable, and that $\mathcal{S}$ is monotone, hence there exists some $\varepsilon > 0$ such that $\boldsymbol{\pi}^n \in \langle \mathcal{S} \rangle$ for every n. Now, for each n let $\boldsymbol{\rho}^n$ be the average of $\boldsymbol{\pi}^m$ over all m such that $\vec{R}_{mn} = 1$, i.e. over all queues at or downstream of n; then $\boldsymbol{\rho}^n \in \langle \mathcal{S} \rangle$ by convexity of $\langle \mathcal{S} \rangle$. Furthermore, we find after a little algebra that $[(I - R^\mathsf{T})\boldsymbol{\rho}^n]_k = \varepsilon 1_{k=n}/(d_n + 1)$ where d_n is the number of queues downstream of n. Finally let $\boldsymbol{\rho}^* = N^{-1}\sum_n \boldsymbol{\rho}^n$; this is in $\langle \mathcal{S} \rangle$ by convexity, and $[(I - R^\mathsf{T})\boldsymbol{\rho}^*]_k = \varepsilon N^{-1}/(d_k + 1) > 0$ for all k.

In each case, both single-hop and multihop, we have found some $\boldsymbol{\rho}^* \in \langle \mathcal{S} \rangle$ such that $(I - R^\mathsf{T})\boldsymbol{\rho}^* > \mathbf{0}$ componentwise. Now let $\mathbf{r}^* = \max(\boldsymbol{\lambda} - (I - R^\mathsf{T})\boldsymbol{\rho}^*, 0)$; clearly $\mathbf{r}^* \in$ FEAS, and by construction either $r^*_n < \lambda_n$ or $\lambda_n = 0$, hence $H(\mathbf{r}^*) < \infty$. The solution to DEP must be at least as good as $\mathbf{r}^*$, so we may as well restrict the domain of DEP to $\mathbf{r} \in$ FEAS such that $H(\mathbf{r}) \leq H(\mathbf{r}^*)$, which implies in particular that $\mathbf{r} \leq \boldsymbol{\lambda}$. Since H is convex, this gives us a closed bounded convex set. And H is strictly convex and finite on it. Hence DEP has a unique solution, call it $\hat{\mathbf{q}}$, and $\hat{\mathbf{q}} \leq \boldsymbol{\lambda}$.

In order to prove that $\hat{q}_n < \lambda_n$ on every queue with $\lambda_n > 0$, i.e. that every queue with arrivals is being served a little, we will first construct an alternative $\mathbf{q} = \hat{\mathbf{q}} - \varepsilon\boldsymbol{\xi}$ such that $\mathbf{q} \in$ FEAS and $q_n < \lambda_n$ on every queue with $\lambda_n > 0$. Since $\hat{\mathbf{q}} \in$ FEAS, we can write $\hat{\mathbf{q}} = \boldsymbol{\lambda} - (I - R^\mathsf{T})\hat{\boldsymbol{\rho}} + \hat{\mathbf{z}}$ for some $\hat{\boldsymbol{\rho}} \in \langle \mathcal{S} \rangle$ and $\hat{\mathbf{z}} \in \mathbb{R}^N_+$. Now pick some small $\varepsilon > 0$, let $\boldsymbol{\rho} = (1 - \varepsilon)\hat{\boldsymbol{\rho}} + \varepsilon\boldsymbol{\rho}^*$, and consider

$$\mathbf{q} = \boldsymbol{\lambda} - (I - R^\mathsf{T})\boldsymbol{\rho} + (1 - \varepsilon)\hat{\mathbf{z}} + \varepsilon\mathbf{z} = \hat{\mathbf{q}} - \varepsilon\big(\hat{\mathbf{z}} - (I - R^\mathsf{T})\hat{\boldsymbol{\rho}} + (I - R^\mathsf{T})\boldsymbol{\rho}^* - \mathbf{z}\big)$$

where

$$z_n = \begin{cases} \big[\big(\hat{\mathbf{z}} - (I - R^\mathsf{T})\hat{\boldsymbol{\rho}} + (I - R^\mathsf{T})\boldsymbol{\rho}^*\big)^+\big]_n & \text{if } \hat{q}_n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Is $\mathbf{q} \in$ FEAS? It is clear that $\mathbf{q} \geq \boldsymbol{\lambda} - (I - R^{\mathsf{T}})\boldsymbol{\rho}$ and $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$. Also, the $\mathbf{z}$ term and a sufficiently small choice of ε guarantee $\mathbf{q} \geq \mathbf{0}$. Also, if $\lambda_n = 0$ then $\hat{q}_n = 0$ because $\hat{\mathbf{q}} \in$ FEAS, hence $[(I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}]_n = \hat{z}_n$, thus $q_n = -\varepsilon([(I - R^{\mathsf{T}})\boldsymbol{\rho}^*]_n - z_n)$; using the definition of z_n and the fact that $(I - R^{\mathsf{T}})\boldsymbol{\rho}^* > \mathbf{0}$ componentwise, we get $q_n = 0$. Thus $\mathbf{q} \in$ FEAS. Furthermore, a similar argument shows that if $\hat{q}_n = \lambda_n > 0$ then $q_n < \lambda_n$, i.e. $\xi_n > 0$.

Now consider how the objective function for DEP changes as we move along the trajectory from $\hat{\mathbf{q}}$ in direction $-\boldsymbol{\xi}$:

$$\frac{dH(\hat{\mathbf{q}} - \varepsilon\boldsymbol{\xi})}{d\varepsilon} = -\sum_{n:\lambda_n>0} w_n \xi_n \Big(\frac{\lambda_n}{\hat{q}_n - \varepsilon\eta_n} - 1\Big)^{-\alpha}$$

Suppose the optimal solution to DEP, $\hat{\mathbf{q}}$, has some queues with $\hat{q}_n = \lambda_n > 0$. Since $\xi_n > 0$ for all these queues, $dH/d\varepsilon = -\infty$ at $\varepsilon = 0$. But we chose $\boldsymbol{\xi}$ so that $\hat{\mathbf{q}} - \varepsilon\boldsymbol{\xi} \in$ FEAS for sufficiently small ε, thus we obtain a contradiction to the optimality of $\hat{\mathbf{q}}$. We conclude that $\hat{q}_n < \lambda_n$ on all queues with $\lambda_n > 0$.

Proof of (ii). Write the objective function for the optimization in (10) as $G((I - R^{\mathsf{T}})\boldsymbol{\rho})$ where

$$G(\boldsymbol{\eta}) = \mathbf{w}\mathbf{q}^{\alpha} \cdot g_{\alpha}(\boldsymbol{\eta}).$$

We found above a $\boldsymbol{\rho}^* \in \langle \mathcal{S} \rangle$ such that $(I - R^{\mathsf{T}})\boldsymbol{\rho}^* > \mathbf{0}$ componentwise, hence $G((I - R^{\mathsf{T}})\boldsymbol{\rho}^*)$ is finite. The solution to the optimization must be at least as good as $\boldsymbol{\rho}^*$, so we might as well restrict the domain to

$$\big\{\boldsymbol{\rho} \in \langle \mathcal{S} \rangle \,:\, (I - R^{\mathsf{T}})\boldsymbol{\rho} \geq \mathbf{0} \text{ and } G(\boldsymbol{\rho}) \geq G(\boldsymbol{\rho}^*)\big\};$$

on this domain the objective is finite. Furthermore the domain is bounded, and it is convex since G is concave. Since we are maximizing a finite concave function over a convex bounded domain, the maximum is attained. □

We now give three lemmas which mirror those used for the max-weight proof. Lemma 5 defines an optimization problem ALGP which is very similar to ALGP for max-weight, and shows that $\hat{\mathbf{q}}$ is its unique soluion (cf. Lemma 1). Unlike with max-weight, this version of ALGP has $\hat{\mathbf{q}}$ appearing in the objective function, so it is not helpful for defining $\hat{\mathbf{q}}$, which is why we used DEP instead. Lemma 6 shows that $\mathbf{q}(t)/t \in$ FEAS for any fluid model solution $\mathbf{q}(\cdot)$ (cf. Lemma 2). Lemma 7 shows that the function $L(\cdot)$ appearing in ALGP is a Lyapunov function (cf. Lemma 3). This function L is closely related to the Lyapunov function introduced by Bonald and Massoulie [1] to prove stability of the bandwidth-sharing model. The difference is that we have added an extra term involving $\hat{q}_n$ to accommodate overload.

Finally, the proof of the main theorem for α-fair, Theorem 2, is identical to the proof of the main theorem for max-weight given in Section 5. The only difference is that appeals to lemmas 1, 2 and 3 should be replaced by appeals to lemmas 5, 6 and 7 respectively. Note that the terms FEAS, ALGP, L and $\hat{\mathbf{q}}$, which appear in the proof of the main theorem, now have different definitions.

It remains to state and prove the three lemmas about α-fair.

Lemma 5 *Given* $\hat{\mathbf{q}}$ *from Definition 5, define* ALGP *to be the following optimization problem:*

$$\text{minimize } L(\mathbf{r}) = \frac{1}{1+\alpha} \sum_{n:\lambda_n>0} w_n r_n^{1+\alpha} (\lambda_n - \hat{q}_n)^{-\alpha} \quad \text{over } \mathbf{r} \in \text{FEAS}\,. \tag{17}$$

(By Lemma 4(i), $\hat{q}_n < \lambda_n$ *for all queues* n *where* $\lambda_n > 0$*, so the objective function makes sense.) Then* ALGP *has a unique solution, which is* $\hat{\mathbf{q}}$*.*

Proof First, observe that that ALGP has a unique minimum. This is because the feasible set is convex and closed; and $L(\mathbf{r}) \to \infty$ as $|\mathbf{r}| \to \infty$, so we may as well restrict the optimization to a bounded subset of FEAS; and L is strictly convex.

The feasible set FEAS is defined by a finite number of linear constraints: that $\mathbf{r} \geq \mathbf{0}$, that $r_n = 0$ for queues n where $\lambda_n = 0$, and that $\mathbf{r} \geq \mathbf{q}$ for some $\mathbf{q}$ in the convex polytope $\{\boldsymbol{\lambda} - (I - R^{\mathsf{T}})\boldsymbol{\rho} : \boldsymbol{\rho} \in \langle \mathcal{S} \rangle\}$; this last polytope has finitely many faces because $\mathcal{S}$ is finite. Now consider the Lagrangian for the optimization problem DEP. At the optimum, there exist dual variables $\hat{\eta}$ such that the complementary slackness conditions are satisfied: for all queues n where $\lambda_n > 0$,

$$w_n \Big(\frac{\lambda_n}{\hat{q}_n} - 1 \Big)^{-\alpha} = \sum_i A_{ni} \hat{\eta}_i$$

where i indexes the constraints of FEAS, and the matrix A indicates which constraints involve which of the q_n. We know that $\hat{q}_n < \lambda_n$, hence the η_i are all finite. Now consider the Lagrangian for the optimization problem ALGP. The complementary slackness conditions are: for all queues n where $\lambda_n > 0$,

$$w_n \Big(\frac{r_n}{\lambda_n - \hat{q}_n} \Big)^{\alpha} = \sum_i A_{ni} \eta_i.$$

The feasible set is the same for ALGP as for DEP, so the constraint matrix A is the same. These conditions are satisfied by setting $r_n = \hat{q}_n$ and $\eta_i = \hat{\eta}_i$. Hence $\hat{\mathbf{q}}$ solves ALGP. □

Lemma 6 *For any fluid model solution* $\mathbf{q}(\cdot)$ *of the* α*-fair policy such that* $q_n(0) = 0$ *for all queues* n *where* $\lambda_n = 0$*,* $\mathbf{q}(t)/t \in \text{FEAS}$ *for all* $t > 0$*.*

Proof Lemma 2 shows that $\mathbf{q}(t)/t$ satisfies the first constraint of FEAS. For the second constraint: by differentiating (2), $\dot{\mathbf{q}}(t) = \boldsymbol{\lambda} - (I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) + \dot{\mathbf{z}}(t)$. The definition of the α-fair policy, (10), says that $(I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) \geq \mathbf{0}$. Thus, for any queue n for which $\lambda_n = 0$, $\dot{q}_n(t) \leq \dot{z}_n(t)$. But whenever $q_n(t) > 0$, equation (5) tells us that $\dot{z}_n(t) = 0$. Since $q_n(\cdot)$ is absolutely continuous, and $\dot{q}_n(t) < 0$ if $q_n(t) > 0$, we conclude $q_n(t) = 0$ for all t. □

Lemma 7 *For the network specified in Definition 5, and with* $L(\cdot)$ *as defined in Lemma 5,*

$$t \frac{d}{dt} L\Big(\frac{\mathbf{q}(t)}{t} \Big) \leq L(\hat{\mathbf{q}}) - L\Big(\frac{\mathbf{q}(t)}{t} \Big) \leq 0.$$

Proof For the first inequality, the drift we want to bound is

$$t \frac{d}{dt} L\Big(\frac{\mathbf{q}(t)}{t} \Big) = \sum_{n:\lambda_n>0} w_n \Big(\frac{q_n(t)/t}{\lambda_n - \hat{q}_n} \Big)^{\alpha} \Big(\frac{dq_n(t)}{dt} - \frac{q_n(t)}{t} \Big).$$

To simplify the notation, we shall (for this proof only) suppress indexes n for which $\lambda_n = 0$ from the dot product. With this convention, the drift we want to bound is

$$\begin{aligned} &= \mathbf{w}\Big(\frac{\mathbf{q}(t)/t}{\boldsymbol{\lambda} - \hat{\mathbf{q}}}\Big)^{\alpha} \cdot \Big(\frac{d\mathbf{q}(t)}{dt} - \frac{\mathbf{q}(t)}{t}\Big) \\ &= \mathbf{w}\Big(\frac{\mathbf{q}(t)/t}{\boldsymbol{\lambda} - \hat{\mathbf{q}}}\Big)^{\alpha} \cdot \Big(\boldsymbol{\lambda} - (I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) - \frac{\mathbf{q}(t)}{t}\Big) \end{aligned} \tag{18}$$

by (2) and (5). We will shortly show that

$$\mathbf{w}\Big(\frac{\mathbf{q}(t)}{\boldsymbol{\lambda} - \hat{\mathbf{q}}}\Big)^{\alpha} \cdot (I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) \geq \mathbf{w}\Big(\frac{\mathbf{q}(t)}{\boldsymbol{\lambda} - \hat{\mathbf{q}}}\Big)^{\alpha} \cdot \big(\boldsymbol{\lambda} - \hat{\mathbf{q}}\big). \tag{19}$$

Multiplying each side by $t^{-\alpha}$ and substituting this into (18),

$$t\frac{d}{dt}L\Big(\frac{\mathbf{q}(t)}{t}\Big) \leq \mathbf{w}\Big(\frac{\mathbf{q}(t)/t}{\boldsymbol{\lambda} - \hat{\mathbf{q}}}\Big)^{\alpha} \cdot \Big(\hat{\mathbf{q}} - \frac{\mathbf{q}(t)}{t}\Big).$$

Finally, by convexity of $L(\cdot)$,

$$\begin{aligned} L(\hat{\mathbf{q}}) &\geq L\Big(\frac{\mathbf{q}(t)}{t}\Big) + \nabla L\Big(\frac{\mathbf{q}(t)}{t}\Big) \cdot \Big(\hat{\mathbf{q}} - \frac{\mathbf{q}(t)}{t}\Big) \\ &= L\Big(\frac{\mathbf{q}(t)}{t}\Big) + \mathbf{w}\Big(\frac{\mathbf{q}(t)/t}{\boldsymbol{\lambda} - \hat{\mathbf{q}}}\Big)^{\alpha} \cdot \Big(\hat{\mathbf{q}} - \frac{\mathbf{q}(t)}{t}\Big) \end{aligned}$$

and rearranging this gives the desired result. It remains to prove (19). We have two separate arguments, one for single-hop and one for multihop networks.

Proof of (19) *for single-hop network.* Define

$$G(\boldsymbol{\eta}) = \mathbf{w}\mathbf{q}(t)^{\alpha} \cdot g_{\alpha}(\boldsymbol{\eta})$$

where g_α was given in the definition of the α-fair policy. The policy chooses $\boldsymbol{\sigma}(t)$ to maximize $G(\boldsymbol{\rho})$ over all $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$. (Note that our notation in this proof suppresses those queues with $\lambda_n = 0$, and by Lemma 6 these queues have $q_n(t) = 0$, and the α-fair policy gives zero weight to empty queues. Thus the special notation in this proof does not cause any problems.)

Observe that G is concave over $\boldsymbol{\eta} \geq \mathbf{0}$; and $\boldsymbol{\sigma}(t) \geq \mathbf{0}$ and $\boldsymbol{\lambda} - \hat{\mathbf{q}} \geq \mathbf{0}$ by Lemma 4(i); hence

$$G\big(\boldsymbol{\sigma}(t)\big) \leq G(\boldsymbol{\lambda} - \hat{\mathbf{q}}) + \nabla G(\boldsymbol{\lambda} - \hat{\mathbf{q}}) \cdot \big(\boldsymbol{\sigma}(t) - (\boldsymbol{\lambda} - \hat{\mathbf{q}})\big).$$

Rearranging, and writing out ∇G explicitly,

$$\mathbf{w}\Big(\frac{\mathbf{q}(t)}{\boldsymbol{\lambda} - \hat{\mathbf{q}}}\Big)^{\alpha} \cdot \big(\boldsymbol{\sigma}(t) - (\boldsymbol{\lambda} - \hat{\mathbf{q}})\big) \geq G\big(\boldsymbol{\sigma}(t)\big) - G(\boldsymbol{\lambda} - \hat{\mathbf{q}}).$$

Since $\hat{\mathbf{q}} \in$ FEAS there exists some $\hat{\boldsymbol{\rho}} \in \langle \mathcal{S} \rangle$ such that $\hat{\mathbf{q}} \geq \boldsymbol{\lambda} - \hat{\boldsymbol{\rho}}$. Since G is increasing componentwise, and since $\boldsymbol{\sigma}(t)$ is chosen to maximize G and $\hat{\boldsymbol{\rho}}$ is a feasible choice,

$$G\big(\boldsymbol{\sigma}(t)\big) - G(\boldsymbol{\lambda} - \hat{\mathbf{q}}) \geq G\big(\boldsymbol{\sigma}(t)\big) - G(\hat{\boldsymbol{\rho}}) \geq 0.$$

This proves (19) for a single-hop network.

Proof of (19) *for multihop network.* We first argue that $\hat{\mathbf{q}} = \boldsymbol{\lambda} - (I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}$ for some $\hat{\boldsymbol{\rho}} \in \langle \mathcal{S} \rangle$. The constraint that $\hat{\mathbf{q}} \in \mathrm{FEAS}$ requires $\hat{\mathbf{q}} \geq \boldsymbol{\lambda} - (I - R^{\mathsf{T}})\boldsymbol{\rho}$ for some $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$, and Lemma 4(i) shows that $\hat{\mathbf{q}} \leq \boldsymbol{\lambda}$. Recalling that $\vec{R} = (I - R^{\mathsf{T}})^{-1}$ is non-negative, $\mathbf{0} \leq \vec{R}(\boldsymbol{\lambda} - \hat{\mathbf{q}}) \leq \boldsymbol{\rho}$. By monotonicity, $\vec{R}(\boldsymbol{\lambda} - \hat{\mathbf{q}}) = \hat{\boldsymbol{\rho}}$ for some $\hat{\boldsymbol{\rho}} \in \langle \mathcal{S} \rangle$, hence $\hat{\mathbf{q}} = \boldsymbol{\lambda} - (I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}$.

Now, consider the choice made by the α-fair policy at time t: it chooses $\boldsymbol{\sigma}(t)$ to maximize $G\big((I - R^{\mathsf{T}})\boldsymbol{\rho}\big)$ over all $\boldsymbol{\rho} \in \langle \mathcal{S} \rangle$ such that $(I - R^{\mathsf{T}})\boldsymbol{\rho} \geq \mathbf{0}$. Our choice $\hat{\boldsymbol{\rho}}$ satisfies this constraint. Using concavity of G and optimality of $\boldsymbol{\sigma}(t)$ as in the single-hop case, we obtain

$$\mathbf{w}\Big(\frac{\mathbf{q}(t)}{(I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}}\Big)^{\alpha} \cdot \big((I - R^{\mathsf{T}})\boldsymbol{\sigma}(t) - (I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}\big) \geq G\big((I - R^{\mathsf{T}})\boldsymbol{\sigma}(t)\big) - G\big((I - R^{\mathsf{T}})\hat{\boldsymbol{\rho}}\big) \geq 0.$$

Rearranging this inequality yields (19) for a multihop network. This completes the proof.

Second inequality. The second inequality in Lemma 7 is straightforward, just as in Lemma 3. □

7 Discussion

In this paper, we have studied fluid models of a switched network in overload, under the max-weight and α-fair algorithms. We have shown that queue sizes grow linearly in time, and characterized the growth rates.

One might ask what the purpose is of studying overload in the way we have. Any real system has limited buffers, so queues will eventually fill up. So what is the relevance in proving, as we have in this paper, that queue sizes grow linearly as time tends to infinity?

One response is to extend the work to incorporate impatient users, who leave the system after a certain random abandonment time. An heuristic analysis has been proposed [6], for α-fair scheduling in the bandwidth-sharing model, but it has not been formally proved.

Another response is to try to use our results to help design discard policies. This might be useful for example in a data center, where queues contain requests and there is no mechanism for customers to remove requests once they have entered the system. But the system does not actually need to keep all unfinished requests in its queues: it could choose to discard requests at rate $\hat{\mathbf{q}}$, where $\hat{\mathbf{q}}$ is the vector of queue growth rates from Theorem 1, while keeping count of all requests (served plus discarded) in a virtual queue, and use the max-weight scheduling rule based on virtual queue sizes rather than real queues. In this way the virtual queues would grow like $t\hat{\mathbf{q}}$, but the actual queues could be kept small. For this scheme to be useful in practice, one needs to strike a balance between letting the virtual queues grow (since this is the learning mechanism by which the network adapts itself to the current arrival rates and averages out random bursts in arrivals), and making the virtual queues 'forget' (in order to adapt quickly when arrival rates change). We are investigating this in an ongoing work.

Finally, there seem to be deep links between fluid limits in overload, and large deviations performance analysis in underload. Specifically, Venkataramanan and Lin [17] have analysed a class of switched networks running the max-weight policy, and obtained a large deviations principle. Their rate function involves the optimization problem from Definition 4. One might expect there to be a link between large deviations and overload, in essence because large deviations asks the question "What is the most likely overload arrival rate that will lead to queue sizes exceeding a certain threshold?" and our analysis asks "Given overload

arrival rates, at what rate do queue sizes grow?" In some sense, if a scheduling algorithm behaves well in overload, then it will require sustained high arrival rates to cause overflow, so the rate function should be large, which suggests that average queue sizes should be small in the stable regime. Much more work is needed before we understand these connections.

Acknowledgements DS is supported by NSF CAREER CNS-0546590. DJW is supported by a Royal Society university research fellowship. We are grateful for further support from the British Council Researcher Exchange program, and the Newton Institute programme on Stochastic Processes in Communication Sciences. We are also grateful to Mike Walfish and Joshua Leners for helpful discussions, and to Adam Greenhalgh for motivating the work.

References

1. Bonald T, Massoulie L (2001) Impact of fairness on internet performance. In: Proceedings of ACM Sigmetrics
2. Chan CW, Armony M, Bambos N (2011) Fairness in overloaded parallel queues, personal communication
3. Dai JG, Lin W (2005) Maximum pressure policies in stochastic processing networks. Operations Research 53(2)
4. Dai JG, Lin W (2008) Asymptotic optimality of maximum pressure policies in stochastic processing networks. The Annals of Applied Probability 18(6)
5. Dai JG, Prabhakar B (2000) The throughput of switches with and without speed-up. In: Proceedings of IEEE Infocom, pp 556–564
6. Egorova R, Borst S, Zwart B (2007) Bandwidth-sharing networks in overload. Performance Evaluation 64:978–993
7. Georgiadis L, Tassiulas L (2006) Optimal overload response in sensor networks. IEEE Transactions on Information Theory 52(6):2684–2696
8. Gromoll HC, Williams RJ (2009) Fluid limits for networks with bandwidth sharing and general document size distributions. The Annals of Applied Probability 19(1)
9. Harrison JM, Zeevi A (2005) A method for staffing large call centers based on stochastic fluid models. Manufacturing & Service Operations Management
10. Jacobson V (1988) Congestion avoidance and control. In: Proceedings of SIGCOMM
11. Kelly FP, Williams RJ (2004) Fluid model for a network operating under a fair bandwidth-sharing policy. The Annals of Applied Probability 14:1055–1083
12. Klemm F, Boudec JYL, Aberer K (2006) Congestion control for distributed hash tables. In: IEEE Symposium on Network Computing and Applications, DOI http://doi.ieeecomputersociety.org/10.1109/NCA.2006.19
13. Mo J, Walrand J (2000) Fair end-to-end windows-based congestion control. IEEE/ACM Transactions on Networking 8(5):556–567
14. Roberts J, Massoulie L (2000) Bandwidth sharing and admission control for elastic traffic. Telecommunication Systems 15:185–201
15. Shah D, Wischik D (2011) Switched networks with maximum weight policies: fluid approximation and multiplicative state space collapse. Annals of Applied Probability (to appear)

16. Tassiulas L, Ephremides A (1992) Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. IEEE Transactions on Automatic Control 37:1936–1948
17. Venkataramanan VJ, Lin X (2007) Structural properties of LDP for queue-length based wireless scheduling algorithms. In: Proceedings of Allerton