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

Resource pooling in congested networks: proportional fairness and product form

F.P. Kelly · L. Massoulié · N.S. Walton

Received: date / Accepted: date

Abstract We review two areas of recent research linking proportional fairness with product form networks. The areas concern, respectively, the heavy traffic and the large deviations limiting regimes for the stationary distribution of a flow model, where the flow model is a stochastic process representing the randomly varying number of document transfers present in a network sharing capacity according to the proportional fairness criterion. In these two regimes we postulate the limiting form of the stationary distribution, by comparison with several variants of the fairness criterion. We outline how product form results can help provide insight into the performance consequences of resource pooling.

Keywords Processor sharing · Multi-path routing · Diffusion approximation · Large deviations

1 Introduction

The processor sharing discipline has been of great interest to queueing theorists since it was first used to model time-shared computer systems (20). The discipline provides the two basic features desired in a time-shared system, namely, rapid service for short jobs, and the appearance of a processor continuously available, albeit a processor of varying capacity. The discipline is also remarkably tractable analytically, a feature it shares with other symmetric queues such as the last-come-first-served queue and Erlang's model of

Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WB, United Kingdom
Thomson Technology Paris Laboratory, Boulogne-Billancourt, France

E-mail: f.p.kelly@statslab.cam.ac.uk
E-mail: laurent.massoulie@thomson.net
E-mail: n.s.walton@statslab.cam.ac.uk

a loss system (1; 7; 14). Thus, for example, a processor sharing queue with Poisson arrivals and independent, arbitrarily distributed service requirements has the property that the mean sojourn time in the queue of an arriving job is proportional to the service requirement of that job, with a constant of proportionality that does not depend upon the overall distribution of service requirements other than through the distribution's mean (8).

In recent decades the interest in processor sharing disciplines has extended to communication networks, where the importance of rapid transfers for short files has been stressed recently by (11). For a system with multiple constrained resources there exist several candidates for the natural generalization of processor sharing, reflecting the ambiguity of what might be meant by fair sharing in the network context. A conveniently parameterized family, that of α-fair rate allocations, was introduced in (25). The parameter α lies in the range $(0, \infty)$, and the cases $\alpha \to 0$, $\alpha = 1$ and $\alpha \to \infty$ correspond respectively to an allocation which achieves maximum throughput, is *proportionally fair* or is *max-min fair* (25; 29).

Max-min is the fairness criterion most commonly discussed for communication networks, but it is not the only possibility. Proportional fairness, in particular, has a claim to be the natural network generalization of processor sharing, with a growing literature showing that it has exact or approximate insensitivity properties (22; 23) and important efficiency and robustness properties (3; 21).

One aim of this paper is to further advance this claim, by reviewing two areas of recent research linking proportional fairness with product form networks. We conjecture the heavy traffic and large deviations behaviour of the stationary distribution of a flow model that describes the randomly varying number of flows present in a network sharing capacity according to the proportional fairness criterion. This flow level model is introduced by Massoulié and Roberts (23). In Sections 5 and 6, we provide support for these conjectures by studying the heavy traffic and large deviations behaviour of networks of processor sharing queues. Theorems 1, 2 and 3 establish a close relationship between networks of processor sharing queues and proportional fairness. Networks of processor sharing queues have a product form stationary distribution, and this suggests product form results may hold for other stochastic systems that more explicitly incorporate proportionally fair optimization. Some limitations are necessary: the topology of the network under study may result in modifications of the conjectured product form. In Section 7 we study grid networks, a class of network with a specific topology for which we can explicitly calculate the limiting stationary distribution for the proportionally fair flow model. In Section 8 we study modified proportional fairness, a variant of proportional fairness. These Sections further refine and motivate our conjectures for proportionally fair flow models. These conjectures are presented in Section 9.

There is currently considerable interest in multi-path routing within the Internet, because of its potential to improve reliability, flexibility and efficiency through *resource pooling* (33). The model of (23) has been generalized by Han et al (12) and Key and Massoulié (19) to allow multi-path routing. A second aim of this paper is to outline how product form results can help provide insight into the performance consequences of resource pooling. In particular, these results suggest an approximation for the mean transfer time of a file in a network operating with multi-path routing under the proportional fairness criterion. The approximation is expressed as a simple sum of terms, one for each resource pool traversed by the file. Under the approximation, the network shares the remarkable property of a processor sharing queue, that the mean transfer time of an arriving file is proportional to the size of the file.

2 Flow models, multipath routing and resource pooling

In this section we introduce our model of flow through a congested network. We begin by defining proportional fairness, in both the uni-path and multi-path setting, following (16). Then we describe the stochastic process which is the focus of this paper: the process was introduced and studied by Massoulié and Roberts (23) as a flow-level model of Internet congestion control, and its generalization to allow multi-path routing has been studied by Han et al (12) and Key and Massoulié (19).

2.1 Fair sharing

Consider a network with a set $\bar{\mathcal{J}}$ of *resources*. Let $\bar{C}_j > 0$ be the *capacity* of resource $j \in \bar{\mathcal{J}}$. Let $\mathcal{R}$ be the set of possible *routes*, and suppose that a unit volume of flow on route r consumes an amount $\bar{a}_{jr} \geq 0$ of resource j for each $j \in \bar{\mathcal{J}}$, where $\sum_{j\in\bar{\mathcal{J}}} \bar{a}_{jr} > 0$ for each $r \in \mathcal{R}$. The simplest case is where we can identify each route r with a non-empty subset of $\bar{\mathcal{J}}$, and where $\bar{a}_{jr} = 1$ if $j \in r$, and $\bar{a}_{jr} = 0$ otherwise. In this case $\bar{A} = (\bar{a}_{jr}, j \in \bar{\mathcal{J}}, r \in \mathcal{R})$ is a $0-1$ incidence matrix. Let n_r be the number of flows on route r. We also let $\bar{J} = |\bar{\mathcal{J}}|$ and $R = |\mathcal{R}|$.

How might the capacities $\bar{C} = (\bar{C}_j, j \in \bar{\mathcal{J}})$ be shared over the routes $\mathcal{R}$? An allocation policy $\Lambda(n) = (\Lambda_r(n), r \in \mathcal{R}) \in \mathbb{R}_+^R$ is called *proportionally fair* if $\forall n \in \mathbb{R}_+^R$, $\Lambda_r^{PF}(n) = 0$ when $n_r = 0$ and $\Lambda^{PF}(n)$ solves[1]

$$\text{maximise} \quad \sum_{r\in\mathcal{R}} n_r \log \Lambda_r \tag{2.1}$$

$$\text{subject to} \quad \sum_{r\in\mathcal{R}} \bar{a}_{jr}\Lambda_r \leq \bar{C}_j, \qquad j \in \bar{\mathcal{J}}, \tag{2.2}$$

$$\text{over} \quad \Lambda_r \geq 0, \qquad r \in \mathcal{R}. \tag{2.3}$$

More generally, an allocation policy is feasible if each allocation $\Lambda(n)$ satisfies constraints (2.2-2.3) $\forall n \in \mathbb{R}_+^R$.

[1] We assume throughout this paper that $xlogx = 0$ for $x = 0$, and we adopt the convention that $0^0 = 1$.

2.2 Multi-path routing and resource pooling

Next we describe a generalization of the earlier model that allows multi-path routing. Let $\mathcal{S}$ be a set of *source-destination pairs* where $s \in \mathcal{S}$ is a non-empty subset of the set of routes $\mathcal{R}$: we interpret $r \in s$ as indicating that the route r is available to carry flow between the source-destination pair s. Let $H_{sr} = 1$ if $r \in s$, and let $H_{sr} = 0$ otherwise. Thus H is an incidence matrix containing only zeros and ones, and $\sum_{s \in \mathcal{S}} H_{sr} = 1$ for each $r \in \mathcal{R}$. Here we consider $n = (n_s : s \in \mathcal{S})$, where we let n_s be the number of flows between source-destination pair s. We also let $S = |\mathcal{S}|$.

In this multi-path setting, an allocation policy $\Lambda(n) = (\Lambda_s(n), s \in \mathcal{S}) \in \mathbb{R}_+^S$ is called *proportionally fair* if $\forall n \in \mathbb{R}_+^S$, $\Lambda_s^{PF}(n) = 0$ when $n_s = 0$ and $\Lambda^{PF}(n)$ solves

$$\begin{aligned} &\text{maximise} && \sum_{s \in \mathcal{S}} n_s \log \Lambda_s && && (2.4)\\ &\text{subject to} && \sum_{r \in \mathcal{R}} \bar{a}_{jr} y_r \le \bar{C}_j, && j \in \bar{\mathcal{J}}, && (2.5)\\ & && \sum_{r \in \mathcal{R}} H_{sr} y_r = \Lambda_s, && s \in \mathcal{S}, && (2.6)\\ &\text{over} && y_r \ge 0,\ r \in \mathcal{R} \quad \text{and} && \Lambda_s \ge 0,\ s \in \mathcal{S}. && (2.7) \end{aligned}$$

In this formulation, the variable y_r represents the flow on route r, and the equation (2.6) expresses the flow between source-destination pair s as the sum of the flows over the routes serving source-destination pair s.

It is possible to rewrite the optimization problem (2.4-2.7) without the variables $y_r, r \in \mathcal{R}$. Consider the problem

$$\begin{aligned} &\text{maximise} && \sum_{s \in \mathcal{S}} n_s \log \Lambda_s && && (2.8)\\ &\text{subject to} && \sum_{s \in \mathcal{S}} a_{js} \Lambda_s \le C_j, && j \in \mathcal{J}, && (2.9)\\ &\text{over} && \Lambda_s \ge 0, && s \in \mathcal{S}. && (2.10) \end{aligned}$$

Then (13, Proposition 5.1) there exists a choice of $\mathcal{J}, A = (a_{js}, j \in \mathcal{J}, s \in \mathcal{S}), C$ independent of n such that C has positive elements, A has non-negative elements and no column of A is identically zero, and such that the unique solution for $(\Lambda_s, s : n_s > 0)$ to the optimization problem (2.8-2.10) is also the unique solution $(\Lambda_s, s : n_s > 0)$ to the optimization problem (2.4-2.7).

The set $\mathcal{J}$ labels a set of virtual resources, or *resource pools* (33). A resource pool might, in a simple case, correspond to a cut set of resources. Typically resource pools correspond to generalizations of cut constraints, and illustrative examples are described in (13; 17; 33). We define $J = |\mathcal{J}|$, the total number of resources.

Note that while as in Section 2.1 it may be natural for the matrix $\bar{A}$ to be a $0 - 1$ route-resource incidence matrix, even in that case the elements of the matrix A, corresponding to resource requirements at pooled resources, may be non-integral. Note also that in the case where the matrix H is the identity matrix, the multi-path routing model reduces to the simpler model of Section 2.1.

The problem (2.8-2.10) is a straightforward convex optimization problem, with optimal solution

$$\Lambda_s^{PF}(n) = \frac{n_s}{\sum_{s\in\mathcal{S}} p_j a_{js}}, \qquad s \in \mathcal{S} \tag{2.11}$$

where the Lagrange multipliers $(p_j, j \in \mathcal{J})$ satisfy

$$p_j \geq 0, \quad p_j(C_j - \sum_{s\in\mathcal{S}} a_{js}\Lambda_s^{PF}(n)) = 0, \quad j \in \mathcal{J}. \tag{2.12}$$

2.3 Flow level model

An allocation $\Lambda(n)$ describes how capacities are shared, for a given number of flows n_s on each source-destination pair $s \in \mathcal{S}$. Next we describe a stochastic model (23; 12; 19) for how the number of flows within the network varies over time.

For an allocation policy $\Lambda : \mathbb{Z}_+^S \to \mathbb{R}_+^S$, define a *$\Lambda$-stochastic flow level model* to be a continuous-time Markov chain on $\mathbb{Z}_+^S = \{0, 1, 2, ...\}^S$ with rates

$$q(n, n') = \begin{cases} \nu_s & \text{if } \ n' = n + e_s, \\ \mu_s \Lambda_s(n) & \text{if } \ n' = n - e_s \text{ and } n_s > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2.13}$$

$\forall n, n' \in \mathbb{Z}_+^S$, where e_s is the s-th unit vector in $\mathbb{Z}_+^S$.

This model can be interpreted as follows. Documents (or files) wishing to be transferred between source-destination pair s arrive as a Poisson process of rate ν_s. These documents are assumed to have a size that is independent and exponentially distributed with mean μ_s^{-1}. If currently the number of documents in transfer across routes is given by the vector $n \in \mathbb{Z}_+^S$ then each document on route s is transferred at rate $\Lambda_s(n)/n_s$. Documents are processed at this rate until there is a change in the network's state, caused either by a document transfer being completed, or by a document arrival occurring.

We can extend the definition of a stochastic flow level model, described in the last paragraph, so that the sizes of incoming documents are independent and of any positive distribution. Information on residual document sizes would be needed for such processes to be Markov. Given this extension, a stochastic flow level model with mean document sizes given by $(\mu_i^{-1} : i \in \mathcal{I})$ is *insensitive* if the stationary distribution for the number of documents in transfer does not depend on the distributions of document size other than through the means $(\mu_i^{-1} : i \in \mathcal{I})$. For more details please refer to Bonald and Proutiere (4; 5).

If the allocation policy is proportionally fair, $\Lambda = \Lambda^{PF}(n)$, then, defining $\rho_s = \frac{\nu_s}{\mu_s}$ $s \in \mathcal{S}$, the stochastic flow level model (2.13) is positive recurrent provided (2; 10)

$$\sum_{s\in\mathcal{S}} a_{js}\rho_s < C_j, \qquad j \in \mathcal{J}. \tag{2.14}$$

An aim of this paper is to understand better the stationary distribution of the flow level model when this condition is satisfied.

3 A network of processor sharing queues

In this section we introduce what we call *a network of processor sharing queues*. Customers in this network belong to different classes and the load different customer classes offer at different queues is given by the entries of the matrix A, from Section 2.2. Thus our queueing network will implicitly share the capacity constraints (2.14). In this section we collect some well known results about the product form stationary distribution of queueing networks.

3.1 Definition

We now more precisely define *a network of processor sharing queues*. We consider a network of queues indexed by the set of resources $\mathcal{J}$. Each queue $j \in \mathcal{J}$ operates under a processor sharing service discipline and has service capacity C_j. Each customer within the network has a class. The set of customer classes is indexed by the set $\mathcal{S}$, the set of source-destination pairs. A customer of class $s \in \mathcal{S}$ at queue $j \in \mathcal{J}$ has an independent exponentially distributed service requirement with mean $\frac{a_{js}}{\mu_s}$.[2] Customers of each class $s \in \mathcal{S}$ arrive into the network as a Poisson process of rate ν_s and we define traffic intensities by the notation $\rho_s = \frac{\nu_s}{\mu_s}$. Upon arrival a customer chooses to visit, independently and with equal probability, a queue from the set $\mathcal{J}$. Similarly a customer which has just completed its service at queue j_k and has visited queues $j_1, ..., j_{k-1} \in \mathcal{J}$ will choose its next queue independently with equal probability from the set $\mathcal{J}\backslash\{j_1, ..., j_k\}$. Once a customer has completed its service requirement at all queues it leaves the network.

3.2 Additional notation

We now introduce some additional notation. The vector $n = (n_s : s \in \mathcal{S}) \in \mathbb{Z}_+^S = \{0, 1, 2, ...\}^S$ will be used to quantify the number of customers of each class in our queueing network and the vector $m = (m_{js} : j \in \mathcal{J}, s \in \mathcal{S}) \in \mathbb{Z}_+^{J\times S}$

[2] If $a_{js} = 0$ we assume a customer instantaneously completes its service, thus in effect never visits the queue. Processor sharing queues are insensitive thus the assumption that the service requirement of customers is exponentially distributed is not necessary. We include this assumption for the convenience of our analysis.

will be used to quantify the number of customers of each class at each queue. Thus we have that

$$n_s = \sum_{j \in \mathcal{J}} m_{js}, \qquad s \in \mathcal{S}.$$

We let $(m_j : j \in \mathcal{J}) \in \mathbb{Z}_+^J$ give the number of customers at each queue, so that

$$m_j = \sum_{s \in \mathcal{S}} m_{js}, \qquad j \in \mathcal{J}.$$

For each $n \in \mathbb{Z}_+^I$ we define $\mathcal{X}(n) = \{m \in \mathbb{Z}_+^{J \times S} : \sum_{j \in \mathcal{J}} m_{js} = n_s, s \in \mathcal{S}\}$, the set of queue states achievable given the number of customer in each class. [3] We also define for each $m \in \mathbb{Z}_+^{J \times S}$

$$\binom{m_j}{m_{js} \;: s \in \mathcal{S}} = \frac{m_j!}{\prod_{s \in \mathcal{S}} (m_{js}!)}.$$

3.3 Stationary distributions

Let $M_{js}(t)$ record the number of class s customers at queue j at time t in a network of processor sharing queues, and let $M(t) = (M_{js}(t), j \in \mathcal{J}, s \in \mathcal{S})$. Note that $(M(t), t \in \mathbb{R}_+)$ is not Markov: but a Markov description could be constructed by augmenting $M(t)$ with information on which queues each customer has already visited. From standard results (1; 7; 8; 14) on product form queueing networks we readily deduce the following proposition and corollaries.

Proposition 1 *A network of processor sharing queues has stationary distribution*

$$\mathbb{P}(M = m) = B^{-1} \prod_{j \in \mathcal{J}} \left(\binom{m_j}{m_{js} \;: s \in \mathcal{S}} \prod_{s \in \mathcal{S}} \left(\frac{a_{js} \rho_s}{C_j} \right)^{m_{js}} \right), \tag{3.1}$$

for each $m \in \mathbb{Z}_+^{J \times S}$, *where*

$$B := \prod_{j \in \mathcal{J}} \left(\frac{C_j}{C_j - \sum_{s \in \mathcal{S}} a_{js} \rho_s} \right), \tag{3.2}$$

provided

$$\sum_{s \in \mathcal{S}} a_{js} \rho_s < C_j, \qquad j \in \mathcal{J}. \tag{3.3}$$

We can also consider the stationary distribution of the number of customers of each class.

[3] In Section 6 when referring to large deviations characteristics n, m and $(m_j : j \in \mathcal{J})$ will be used to refer to proportions of customers within the network.

Corollary 1 *The number of customer in each class, $N = (N_s : s \in \mathcal{S})$, has stationary distribution*

$$\mathbb{P}(N = n) = \frac{B_n}{B} \prod_{s \in \mathcal{S}} \rho_s^{n_s}, \qquad n \in \mathbb{Z}_+^S, \tag{3.4}$$

where

$$B_n := \sum_{m \in \mathcal{X}(n)} \prod_{j \in \mathcal{J}} \left(\binom{m_j}{m_{js} : s \in \mathcal{S}} \prod_{s \in \mathcal{S}} \left(\frac{a_{js}}{C_j} \right)^{m_{js}} \right), \qquad n \in \mathbb{Z}_+^S. \tag{3.5}$$

From this we will be interested in the stationary distribution of a network of processor sharing queues conditional on the number of customers of each class within the network being given by a fixed vector $n = (n_s : s \in \mathcal{S}) \in \mathbb{Z}_+^S$. From the last two results we can deduce that the conditional distribution is given by

$$\mathbb{P}(M = m | N = n) = B_n^{-1} \prod_{j \in \mathcal{J}} \left(\binom{m_j}{m_{js} : s \in \mathcal{S}} \prod_{s \in \mathcal{S}} \left(\frac{a_{js}}{C_j} \right)^{m_{js}} \right) \tag{3.6}$$

for all $m \in \mathcal{X}(n)$. We will be specifically interested in the rate class s customers are processed by a network of processor sharing queues given that the number of customers of each class is equal to n.

Corollary 2 *For a queue $j \in \mathcal{J}$ and a class $s \in \mathcal{S}$ with $a_{js} > 0$, and conditional on there being $n \in \mathbb{Z}_+^S$ customers of each class, the rate class s customers are processed through queue j in a network of processor sharing queues is given by*

$$\mu_s \frac{B_{n-e_s}}{B_n},$$

where B_n is defined by (3.5) and e_s is the s-th unit vector in $\mathbb{Z}_+^S$.

Proof The probability the network is in state $m \in \mathbb{Z}_+^{J \times S}$ is given by (3.6). Thus the throughput of class s customers at queue j is

$$\sum_{\substack{m \in \mathcal{X}(n): \\ m_j > 0}} \frac{C_j m_{js} \mu_s}{a_{js} m_j} \frac{1}{B_n} \prod_{l \in \mathcal{J}} \left(\binom{m_l}{m_{ls'} : s' \in \mathcal{S}} \prod_{s' \in \mathcal{S}} \left(\frac{a_{ls'}}{C_l} \right)^{m_{ls'}} \right)$$

$$= \sum_{m' \in \mathcal{X}(n-e_s)} \frac{\mu_s}{B_n} \prod_{l \in \mathcal{J}} \left(\binom{m'_l}{m'_{ls'} : s' \in \mathcal{S}} \prod_{s' \in \mathcal{S}} \left(\frac{a_{ls'}}{C_l} \right)^{m'_{ls'}} \right) = \mu_s \frac{B_{n-e_s}}{B_n}.$$

Above we cancelled terms and substituted $m'_{ls'} = m_{ls'} - 1$ if $(l, s') = (j, s)$ and $m'_{ls'} = m_{ls'}$ otherwise. □

4 The spinning network

We next define a stochastic flow level model motivated by the network of processor sharing queues considered in the last section. In Corollary 2 we determined the throughput of customers passing through a network of processor sharing queues condition on the number of customers in each class. From this, for $n \in \mathbb{Z}_+^S$, we define the *spinning allocation* to be the allocation policy $\Lambda^{SN}(n) = (\Lambda_s^{SN}(n) : s \in \mathcal{S})$ where

$$\Lambda_s^{SN}(n) = \begin{cases} \frac{B_{n-e_s}}{B_n} & \textit{if } n_s > 0, \\ 0 & \textit{otherwise}, \end{cases} \qquad s \in \mathcal{S}, \tag{4.1}$$

where B_n is defined by (3.5). We call the stochastic flow level model defined by (2.13) and operating under the spinning allocation $\Lambda = \Lambda^{SN}$, the *spinning network*. The spinning network is essentially the flow level generalization of a network of processor sharing queues.

An allocation policy of this type was first considered by Massoulié, and was consequently discussed in the thesis of Proutière (27, Section 3.4). Bonald and Proutière (5; 27) showed that the spinning network is insensitive to different document size distributions. Walton (30) has established the weak convergence of a sequence of processor sharing queueing networks to the spinning network.

Balanced fairness is a further allocation policy that has received attention (5; 27). Balanced fairness has the unique property of being both insensitive and Pareto efficient amongst the set of feasible allocation policies. As we shall discuss in Theorem 3, in addition to being insensitive, the spinning allocations asymptotically approaches the set of Pareto efficient allocations. In particular, we shall see it converges to a proportionally fair allocation (30). It has been conjectured that the balanced fair policy convergences to the proportional fair policy, in the sense described in Theorem 3, see Massoulié (22).

The spinning network is reversible. By checking the detailed balance conditions for the spinning network one can verify that its stationary distribution is

$$\mathbb{P}(N = n) = \frac{B_n}{B} \prod_{s \in \mathcal{S}} \rho_s^{n_s}, \qquad n \in \mathbb{Z}_+^S.$$

From Corollary 1, this is also the stationary distribution for the number of customers of each class in a network of processor sharing queues, which is as we would expect, given the motivation for the definition of the spinning network.

5 Heavy traffic and product form

In this section we consider a network of processor sharing queues as the system approaches overload. We consider a sequence of networks, each in equilibrium, and explore the connection in heavy traffic between proportionally fair stochastic flow level models and product form queueing networks.

For each $h \in \mathbb{N} = \{1, 2, ...\}$, let $M^{(h)}$ be a stationary network of processor sharing queues of the form described in Section 3, with traffic intensities $\rho_s^{(h)} = \rho_s - \frac{\sigma_s}{h}$ for $s \in \mathcal{S}$. We assume that $\sigma_s > 0$ for $s \in \mathcal{S}$. We also assume that $\rho \in \mathbb{R}_+^S = [0, \infty)^S$ is *Pareto efficient*, that is that $A\rho \le C$ and $\forall \delta \in \mathbb{R}_+^S$, $A(\rho + \delta) \le C$ implies $\delta = 0$. We shall say that queue j is in heavy traffic if $(A\rho)_j = C_j$. Thus $M^{(h)}$ has a stationary distribution given by Proposition 1 with its consequence Corollary 1. Thus by definition we have that

$$N_s^{(h)} = \sum_{j \in \mathcal{J}} M_{js}^{(h)}, \qquad s \in \mathcal{S}, \tag{5.1}$$

where for all $j \in \mathcal{J}$ and $s \in \mathcal{S}$

$$M_{js}^{(h)} | M_j^{(h)} \sim \text{Binomial}(M_j, \frac{a_{js}\rho_s^{(h)}}{C_j}) \tag{5.2}$$

and $M_j^{(h)}$, $j \in \mathcal{J}$, are independent with

$$M_j^{(h)} \sim \text{Geometric}(h \frac{\sum_{j \in \mathcal{J}} a_{js}\sigma_s}{C_j}).$$

Letting $h \to \infty$,

$$(\frac{M_j^{(h)}}{h} : j \in \mathcal{J}) \Rightarrow (\hat{M}_j : j \in \mathcal{J})$$

where $\hat{M}_j$ is exponentially distributed with parameter $\frac{\sum_{s \in \mathcal{S}} a_{js}\sigma_s}{C_j}$ if queue j is in heavy traffic, $\hat{M}_j = 0$ otherwise, and $\hat{M}_j, j \in \mathcal{J}$, are independent. By the strong law of large numbers for the binomial distribution (5.2), as $h \to \infty$,

$$(\frac{M_{js}^{(h)}}{h} : j \in \mathcal{J}, s \in \mathcal{S}) \Rightarrow (a_{js}\rho_s \frac{\hat{M}_j}{C_j} : j \in \mathcal{J}, s \in \mathcal{S}).$$

Thus we have that, $\frac{N^{(h)}}{h} \Rightarrow \hat{N}$, where

$$\hat{N}_s = \sum_{j \in \mathcal{J}} a_{js}\rho_s \frac{\hat{M}_j}{C_j}, \qquad \forall s \in \mathcal{S}. \tag{5.3}$$

Thus the (scaled) number of customers on route s is distributed as the sum of independent exponential random variables. In addition by comparison with (5.3) and our conditions on the positivity of M_j, $j \in \mathcal{J}$, we can see that the conditions (2.11-2.12) are satisfied and thus, almost surely, $\Lambda_s^{PF}(\hat{N}) = \rho_s$, $s \in \mathcal{S}$.

Let $\mathcal{J}^* = \{j \in \mathcal{J} : \sum_{s \in \mathcal{S}} a_{js}\rho_s = C_j\}$. So that each route s traverses a queue in heavy traffic we assume no column of the matrix $(a_{js}, j \in \mathcal{J}^*, s \in \mathcal{S})$ is identically zero. We also let $J^* = |\mathcal{J}^*|$. Our argument above can be formalized to give the next theorem.

Theorem 1 *Let $M^{(h)}$ have the stationary distribution (3.1) of a network of processor sharing queues with parameters $\rho^{(h)} = (\rho_s - \frac{\sigma_s}{h} : s \in \mathcal{S})$, and let $N^{(h)}$ be given by (5.1). Then as $h \to \infty$*

$$\left(\frac{M^{(h)}}{h}, \frac{N^{(h)}}{h}\right) \Rightarrow (\hat{M}, \hat{N})$$

where $\hat{M}_j, j \in \mathcal{J}$, are independent, $\hat{M}_j$ is exponentially distributed with parameter $\sum_{s\in\mathcal{S}} \frac{a_{js}\sigma_s}{C_j}$, $j \in \mathcal{J}^*$, and $\hat{M}_j = 0$, $j \in \mathcal{J}\backslash\mathcal{J}^*$, and where $\hat{N}$ is defined by (5.3).

Moreover, almost surely the pair $(p, n) = ((\frac{\hat{M}_j}{C_j})_{j\in\mathcal{J}}, (\hat{N}_s)_{s\in\mathcal{S}})$ satisfies the proportional fairness optimality conditions (2.11-2.12) and

$$\Lambda_s^{PF}(\hat{N}) = \rho_s, \qquad s \in \mathcal{S}.$$

The proof of this theorem can be found in the Appendix A.1.

A connection between multi-class networks of single server queues and the optimization formulation (2.8-2.10) has been noted several times in the literature (15; 24; 28), and the above theorem provides a formalization, in heavy traffic, of the connection.

The support of $\hat{N}$ is the manifold

$$\begin{aligned}\mathcal{N} &= \{n \in \mathbb{R}_+^S : \Lambda_s(n) = \rho_s,\ s \in \mathcal{S}\} \\ &= \{n : \exists q \in \mathbb{R}_+^{\mathcal{J}^*} \text{ s.t. } n_s = \sum_{j\in\mathcal{J}^*} q_j A_{js}\rho_s, \text{ for } s \in \mathcal{S}\}. \end{aligned} \tag{5.4}$$

The second equality above can be deduced from expressions (2.11-2.12) or alternatively seen in (18, Theorem 5.1). In (18) it is shown that the proportionally fair stochastic flow level model (2.13) has a fluid model in heavy traffic which converges to exactly the manifold (5.4). Observe that there is a form of state space collapse: the dimension of the manifold (5.4) is the row rank of the matrix $A_{\mathcal{J}^*} = (a_{js} : j \in \mathcal{J}^*, s \in \mathcal{S})$, at most J^*. In (13) a diffusion approximation is established for the proportionally fair stochastic flow level model, under certain additional conditions, and the stationary distribution for the diffusion approximation matches the distribution for $\hat{N}$ found above. These additional conditions are that $\mathcal{J}^* = \mathcal{J}$, so that all resources are in heavy traffic, and a *local traffic condition*. This condition requires that amongst the columns of the matrix A there are the columns of a diagonal matrix, so that for each resource $j \in \mathcal{J}$ there is a traffic stream $s \in \mathcal{S}$ which uses just resource j.

The local traffic condition implies that A has full row rank, and that the dimension of the manifold $\mathcal{N}$ is $\mathcal{J}^*$. In Section 7 we shall discuss an example where the matrix A is not of full row rank, and where, in heavy traffic, there is a distinction between the limiting stationary distributions of the proportionally fair stochastic flow level model and the corresponding product form processor sharing queueing network.

6 Large deviations and convergence of throughput

In the previous section we considered the stationary behaviour of a network of processor sharing queues in a heavy traffic regime. We found in this analysis that a network of processor sharing queues was able to capture certain aspects related to the multi-path proportionally fair optimization problem.

In this section we continue to pursue this relationship. We now consider a network of processor sharing queues in a large deviations regime. We will show in Theorem 2 that for a network of processor sharing queues the stationary distribution of the number of customers in each class obeys a large deviation principle with good rate function

$$\alpha_\rho(n) = \max_{\Lambda \in \mathbb{R}_+^S} \sum_{s:n_s>0} n_s \log \frac{\Lambda_s}{\rho_s} \quad \text{subject to} \quad \sum_{s\in\mathcal{S}} a_{js}\Lambda_s \le C_j, \quad j \in \mathcal{J}. \tag{6.1}$$

Note that, apart from a constant term added to the objective function, the optimization problem (6.1) is identical to the earlier problem (2.8-2.10).

We also state but do not prove an additional theorem, Theorem 3. Theorem 3 considers the throughput of a network of processor sharing queues conditional on the number of customers of each class being large but proportional to some fixed vector $n \in \mathbb{R}_+^S$. Theorem 3 demonstrates that this quantity converges to a solution of the multi-path proportionally fair optimization problem (2.8-2.10).

The results in this section are proven in (30), with the main distinction being that the paper (30) only allows a customer to have exponentially distributed mean 1 service requirement at each queue it visits. Here we allow customers to have service requirements given by values in the matrix A; this is important in the context of multi-path routing. The large deviations properties of the stationary distributions of product form queueing networks were first considered by Pittel (26). Massoulié (22) first established the rate function (6.1) as the large deviations limit of stochastic flow level models operating under modified proportional fairness.

To prove Theorem 2, first we prove a large deviation principle for the stationary distribution of a network of processor sharing queues (3.1). Stirling's formula is used to find a rate function: label the rate function $\beta_\rho(\cdot)$. Applying the contraction principle gives the large deviation principle for the number of customers in each class and finds $\alpha_\rho(\cdot)$ expressed as the primal form of a convex optimization problem. We calculate the dual of this optimization problem and find it to be of the form of (6.1).

We start by finding the rate function $\beta_\rho(\cdot)$.

Lemma 1 *Suppose random variable M in $\mathbb{Z}_+^{J\times S}$ has the stationary distribution of the number of customers of each class at queues in a network of processor sharing queues (3.1). If we take a vector $m \in \mathbb{R}_+^{J\times S}$ and take $\{d^{(h)}\}_{h\in\mathbb{N}}$ a sequence of vectors in $\mathbb{R}^{J\times S}$ such that $hm + d^{(h)} \in \mathbb{Z}_+^{J\times S}$ and $\sup_h |d^{(h)}| < \infty$*

then

$$\lim_{h\to\infty}\frac{1}{h}\log\mathbb{P}(M=hm+d^{(h)})=-\beta_\rho(m),$$

where we define

$$\beta_\rho(m):=\sum_{\substack{j\in\mathcal{J},s\in\mathcal{S}:\\ m_j>0,\ a_{js}>0}} m_{js}\log\frac{m_{js}C_j}{m_j a_{js}\rho_s}. \tag{6.2}$$

Proof For all $j\in\mathcal{J}$ define, $d_j^{(h)}=\sum_{s\in\mathcal{S}}d_{js}^{(h)}$. By Stirling's formula

$$\begin{aligned}
&\lim_{h\to\infty}\frac{1}{h}\log\mathbb{P}(M=hm+d^{(h)})\\
&=\lim_{h\to\infty}\frac{1}{h}\Bigg[\sum_{j\in\mathcal{J}}\log\,(hm_j+d_j^{(h)})!-\sum_{j\in\mathcal{J},s\in\mathcal{S}}\log\,(hm_{js}+d_{js}^{(h)})!\\
&\qquad\qquad+\sum_{j\in\mathcal{J},s\in\mathcal{S}}(hm_{js}+d_{js}^{(h)})\log\frac{a_{js}\rho_s}{C_j}\Bigg]\\
&=\lim_{h\to\infty}\frac{1}{h}\Bigg[\sum_{\substack{j\in\mathcal{J}:\\ m_j>0}}\Big((hm_j+d_j^{(h)})\log(hm_j+d_j^{(h)})-(hm_j+d_j^{(h)})\Big)\\
&\qquad-\sum_{\substack{j\in\mathcal{J},s\in\mathcal{S}:\\ m_{js}>0}}\Big((hm_{js}+d_{js}^{(h)})\log(hm_{js}+d_{js}^{(h)})-(hm_{js}+d_{ji}^{(h)})\Big)\\
&\qquad\qquad+\sum_{j\in\mathcal{J},s\in\mathcal{S}}(hm_{js}+d_{js}^{(h)})\log\frac{a_{js}\rho_s}{C_j}\Bigg]\\
&=-\lim_{h\to\infty}\sum_{\substack{j\in\mathcal{J},s\in\mathcal{S}:\\ m_{js}>0}}m_{js}\log\frac{(m_{js}+\frac{d_{js}^{(h)}}{h})C_j}{(m_j+\frac{d_j^{(h)}}{h})a_{js}\rho_s}=-\beta_\rho(m).
\end{aligned}$$

□

From this result one can more formally establish the following large deviation principle. For details of how to formalize the following proposition please see (30, Section 6).

Proposition 2 *If random variable M in $\mathbb{Z}_+^{J\times S}$ has the stationary distribution of the number of customers of each class at queues in a network of processor sharing queues (3.1) then, as $h\to\infty$, $\{\frac{M}{h}\}_{h\in\mathbb{N}}$ obeys a large deviation principle on $\mathbb{R}_+^{J\times S}$ with convex, continuous, good rate function $\beta_\rho(\cdot)$. That is for all $D\subset\mathbb{R}_+^{J\times S}$ Borel measurable*

$$-\inf_{m\in D^\circ}\beta_\rho(m)\le\liminf_{h\to\infty}\mathbb{P}(\frac{M}{h}\in D)\le\limsup_{h\to\infty}\mathbb{P}(\frac{M}{h}\in D)\le-\inf_{m\in\bar{D}}\beta_\rho(m),$$

where D° is the interior of D and $\bar{D}$ is the closure of D.

To prove Theorem 2 we require two technical lemmas about the function $\beta_\rho(\cdot)$.

Lemma 2 *For all* $\Lambda \in (0,\infty)^S$

$$\inf_{m\in\mathbb{R}_+^{J\times S}} \beta_\Lambda(m) = \begin{cases} 0 & \text{if } \sum_{s\in\mathcal{S}} a_{js}\Lambda_s \le C_j, \ \forall j \in \mathcal{J} \\ -\infty & \text{otherwise.} \end{cases}$$

Proof Consider two probability distributions p and q with the same support on $\mathcal{S}$. One can verify with calculus that the relative entropy $D(p||q) = \sum_s p_s \log \frac{p_s}{q_s}$ of the two probability distributions p and q is such that

$$\min_p D(p||q) = 0$$

and is minimized by $p = q$. Thus,

$$\begin{aligned}\inf_{m\in\mathbb{R}_+^{J\times S}} \beta_\Lambda(m) &= \inf_{m\in\mathbb{R}_+^{J\times S}} \sum_{j:m_j>0} m_j \sum_{s\in\mathcal{S}} \frac{m_{js}}{m_j} \log \frac{m_{js}C_j}{m_j a_{js}\Lambda_s} \\ &= \inf_{m'\in\mathbb{R}_+^{J}} \sum_{j:m'_j>0} m'_j \log \frac{C_j}{\sum_{s\in\mathcal{S}} a_{js}\Lambda_s} = \begin{cases} 0 & \text{if } \sum_{s\in\mathcal{S}} a_{js}\Lambda_s \le C_j, \\ -\infty & \text{otherwise.} \end{cases}\end{aligned}$$

□

Lemma 3 *For all* $\Lambda \in (0,\infty)^I$, $\beta_\Lambda(\cdot)$ *is a convex continuous function.*

See (30, Lemma 6.3) for a proof of this lemma.

We now prove Theorem 2. The theorem helps explain how the collapse of our original queueing model is related to the multi-path proportionally fair optimisation problem. On the one hand the expression (6.3) can be interpreted as saying that a network of processor sharing queues wishes to minimize the entropy of queue sizes subject to constraints on the number of customers in each class. On the other hand the dual (6.4) can be interpreted as saying, given the number of customers of each class, flows wish to maximize the proportionally fair optimization problem.

Theorem 2 *If* N *is a random variable in* $\mathbb{Z}_+^S$ *has the stationary distribution of the number of customers of each class in a network of processor sharing queues (3.4) then, as* $h \to \infty$, $\{\frac{N}{h}\}_{h\in\mathbb{N}}$ *obeys a large deviation principle on* $\mathbb{R}_+^S$ *with good rate function*

$$\alpha_\rho(n) := \min_{m\in\mathbb{R}_+^{J\times S}} \sum_{\substack{j\in\mathcal{J}, s\in\mathcal{S}: \\ m_j>0,\ a_{js}>0}} m_{js} \log \frac{m_{js}C_j}{m_j a_{js}\rho_s} \quad \text{subject to} \sum_{j\in\mathcal{J}} m_{js} = n_s,\ s\in\mathcal{S} \tag{6.3}$$

$$= \max_{\Lambda\in\mathbb{R}_+^S} \sum_{s\in\mathcal{S}} n_s \log \frac{\Lambda_s}{\rho_s} \quad \text{subject to} \sum_{s\in\mathcal{S}} a_{js}\Lambda_s \le C_j,\ j\in\mathcal{J}. \tag{6.4}$$

That is, for all Borel measurable $D \subset \mathbb{R}_+^S$ we have that

$$-\inf_{n\in D^\circ} \alpha_\rho(n) \le \liminf_{h\to\infty} \mathbb{P}(\frac{N}{h} \in D) \le \limsup_{h\to\infty} \mathbb{P}(\frac{N}{h} \in D) \le -\inf_{n\in \bar{D}} \alpha_\rho(n).$$

Proof Applying the contraction principle (9, page 126) to Proposition 2 using the continuous map $f : \mathbb{R}_+^{J\times S} \to \mathbb{R}_+^S$ such that $f(m) = (\sum_{j\in\mathcal{J}} m_{js} : s \in \mathcal{S})$ shows that $\{\frac{N}{h}\}_{h\in\mathbb{N}}$ obeys a large deviation principle with good rate function

$$\alpha_\rho(n) = \min_{m\in\mathbb{R}_+^{J\times S}} \sum_{\substack{j\in\mathcal{J},s\in\mathcal{S}:\\ m_j>0}} m_{js} \log \frac{m_{js}C_j}{m_j a_{js}\rho_s} \quad \text{subject to} \quad \sum_{j\in\mathcal{J}} m_{js} = n_s, \quad s\in\mathcal{S}.$$

As β_ρ is convex, this is a convex optimisation problem. Let us calculate its dual formulation. Using Lagrange multipliers $\lambda \in \mathbb{R}^S$, its Lagrangian is

$$\begin{aligned} L(m,\lambda) &= \sum_{\substack{j\in\mathcal{J},s\in\mathcal{S}:\\ m_j>0,n_s>0}} m_{js} \log \frac{m_{js}C_j}{m_j a_{js}\rho_s} + \sum_{s:n_s>0} \lambda_s \left(n_s - \sum_{j\in\mathcal{J}} m_{js} \right) \\ &= \sum_{\substack{j\in\mathcal{J},s\in\mathcal{S}:\\ m_j>0,n_s>0}} m_{js} \log \frac{m_{js}C_j}{m_j a_{js}\rho_s e^{\lambda_s}} + \sum_{s:n_s>0} \lambda_s n_s. \end{aligned}$$

By Lemma 2

$$\min_{m\in\mathbb{R}_+^{J\times S}} L(m,\lambda) = \begin{cases} \sum_{s:n_s>0} n_s\lambda_s & \text{if} \quad \sum_{s\in\mathcal{S}} a_{js}\rho_s e^{\lambda_s} \le C_j, \;\; j\in\mathcal{J}, \\ -\infty & \text{otherwise.} \end{cases}$$

Thus we find its dual is

$$\alpha_\rho(n) = \max_{\lambda\in\mathbb{R}^S} \sum_{s:n_s>0} n_s\lambda_s \quad \text{subject to} \quad \sum_{s\in\mathcal{S}} a_{js}\rho_s e^{\lambda_s} \le C_j, \quad j\in\mathcal{J}.$$

Substituting $\Lambda_s = \rho_s e^{\lambda_s}$ gives

$$\alpha_\rho(n) = \max_{\Lambda\in\mathbb{R}_+^S} \sum_{s:n_s>0} n_s \log \frac{\Lambda_s}{\rho_s} \quad \text{subject to} \quad \sum_{s\in\mathcal{S}} a_{js}\Lambda_s \le C_j, \quad j\in\mathcal{J}.$$

□

An important consequence of this result is given in the next result, Theorem 3. We state the result here, but refer the reader to (24, Section C) for an accessible justification of the result and to (30, Section 7) for a proof. In Theorem 3 we consider $\Lambda_s^{SN}(\lfloor hn \rfloor)$, the throughput of class s customers in a network of processor sharing queues conditional on the number of customers of each class being given by $\lfloor hn \rfloor = (\lfloor hn_s \rfloor : s\in\mathcal{S})$. Theorem 3 states that as $h\to\infty$ this throughput converges to a proportionally fair bandwidth allocation.

Theorem 3 *For all $n \in \mathbb{R}_+^S$ and $s \in \mathcal{S}$*

$$\Lambda_s^{SN}(\lfloor hn \rfloor) \xrightarrow[h\to\infty]{} \Lambda_s^{PF}(n),$$

where we define $\lfloor n \rfloor = (\lfloor n_s \rfloor : s \in \mathcal{S})$.

We note that this connection between the stationary behaviour of a product form queueing network and a multi-path proportionally fair stochastic flow level model has not required any rank condition on the matrix A.

7 Grid networks

In Section 5 we studied the stationary distribution of processor sharing queueing networks in heavy traffic. For grid networks the stationary distribution of proportionally fair stochastic flow level models is known. Therefore we can form an analogous heavy traffic analysis to Section 5. We know from Theorem 3 that asymptotically the throughput of network of processor sharing queues converges to a proportionally fair allocation. Therefore we might expect the two models to agree in heavy traffic. In fact, despite this, we find that under a heavy traffic scaling, the limit distributions of these two models do not agree.

A $K \times L$ grid network is a network with uni-path routing, that is the set of routes can be identified with the set of source-sink pairs. A $K \times L$ grid network has links $\mathcal{J} = \{(k,l) : k = 1, ..., K, l = 1, ..., L\}$ and routes $\mathcal{R} = \mathcal{S} = \{k\}_{k=1}^K \cup \{l\}_{l=1}^L$ where $k = \{(k,l) : l = 1, ..., L\}$ and $l = \{(k,l) : k = 1, ..., K\}$. We refer to routes indexed by k as vertical routes and routes indexed by l as horizontal routes. We let n_{xk} denote the number of horizontal flows on route k and we let n_{yl} denote the number of vertical flows on route l. In addition we use the shorthand n_x and n_y to denote the total number of horizontal flows and vertical flows respectively. We assume all capacities are equal to 1, as are all non-zero entries of the matrix A. For the proportionally fair stochastic flow level model we assume that documents arrive as a Poisson process, of rate ν_x for each vertical route and of rate ν_y for each horizontal route and that document sizes are independent and exponentially distributed, with parameter μ_x for each vertical route and of rate μ_y for each horizontal route. Finally we define $\rho_x = \frac{\nu_x}{\mu_x}$ and $\rho_y = \frac{\nu_y}{\mu_y}$.

We can calculate the rates and stationary distribution of a proportionally fair stochastic flow level model on a grid network. The following proposition is due to Bonald and Massoulié (2).

Proposition 3 *For all $n \in \mathbb{Z}_+^{K+L}$, a $K \times L$ grid network operating under proportional fairness has an allocation*

$$\Lambda_k^{PF}(n) = \frac{\sum_{k=1}^K n_{xk}}{\sum_{k=1}^K n_{xk} + \sum_{l=1}^L n_{yl}}, \quad k = 1, ..., K$$

$$\Lambda_l^{PF}(n) = \frac{\sum_{l=1}^L n_{yl}}{\sum_{k=1}^K n_{xk} + \sum_{l=1}^L n_{yl}}, \quad l = 1, ..., L. \tag{7.1}$$

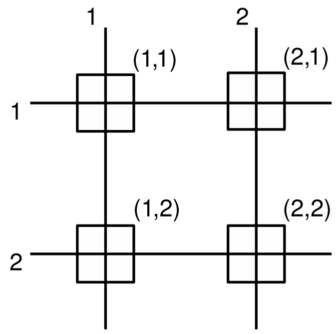

Fig. 1 2×2 grid network.

Its proportionally fair stochastic flow level model has stationary distribution,

$$\mathbb{P}(N = n) = \frac{1}{C(\rho)} \binom{\sum_{k=1}^{K} n_{xk} + \sum_{l=1}^{L} n_{yl}}{\sum_{k=1}^{K} n_{xk}} \rho_x^{\sum_{k=1}^{K} n_{xk}} \rho_y^{\sum_{l=1}^{L} n_{yl}}, \tag{7.2}$$

for $n \in \mathbb{Z}_+^{K+L}$ *where*

$$C(\rho_x, \rho_y) = \sum_{k=1}^{K-1} \sum_{l=1}^{L-1} \binom{k+l}{k} \frac{\rho_x^k \rho_y^l}{(1 - \rho_x - \rho_y)^{k+l+1}}. \tag{7.3}$$

Proof We first confirm (7.1). Note given we hold all vertical rates fixed then the optimal horizontal rates must satisfy $\Lambda_k = 1 - \max_l \Lambda_l \ \forall k = 1, ..., K$. So the optimal choice of horizontal rates has all horizontal rates equal. By symmetry the same must hold for all vertical routes: $\Lambda_l = \Lambda_{l'} \ \forall l, l' = 1, ..., L$. Also the capacity constraint gives that $\Lambda_k = 1 - \Lambda_l$. These equalities reduce the proportionally fair optimisation problem to a problem in two variables which can be solved to give (7.1).

The detailed balance conditions can be checked to show that the process n is reversible, with stationary distribution (7.2).

Finally, the normalizing constant (7.3) is found in Lemma 5 in the Appendix. □

Consider a network of processor sharing queues from Section 3, with the topology of a $K \times L$ grid network. We know, as shown in Theorem 1, that the geometrically distributed queue sizes approach exponential distributions under a heavy traffic scaling. Thus the total number of customers in the processor sharing queueing network, under the same scaling, will approach an Erlang distribution with parameters $K \times L$ and $\sigma_x + \sigma_y$. If the heavy traffic stationary distributions of both the queueing model and the proportionally fair model were the same then the distribution of the total number of documents in transfer in the proportionally fair model would be Erlang$(K \times L, \sigma_x + \sigma_y)$.

The next resultshows that this is not the case and that the distribution is in fact Erlang$(K+L-1,\sigma_x+\sigma_y)$.

Theorem 4 *For each* $h \in \mathbb{N}$*, let* $N^{(h)}$ *have the stationary distribution of a proportionally fair stochastic flow level model on a* $K \times L$ *grid network with traffic intensities* $\rho_x^{(h)} = \rho_x - \frac{\sigma_x}{h}$*,* $\rho_y^{(h)} = \rho_y - \frac{\sigma_y}{h}$ *and* $\rho_x + \rho_y = 1$*. Let* $N'^{(h)}$ *be the total number of documents in transfer in this model, then*

$$\frac{N'^{(h)}}{h} \Rightarrow \hat{N}'$$

where $\hat{N}'$ *has an Erlang distribution with parameters* $K+L-1$ *and* $\sigma_x+\sigma_y$*.*

Proof The moment generating function of distribution of the total number of documents in a $K \times L$ grid network (7.3) is given by,

$$\frac{C(\rho_x e^\theta, \rho_y e^\theta)}{C(\rho)}, \qquad \theta \in \mathbb{C}.$$

The highest order term in (7.3) is from $k = K-1$ and $l = L-1$, in that

$$C(\rho_x^{(h)} e^{\frac{\theta}{h}}, \rho_y^{(h)} e^{\frac{\theta}{h}}) = \frac{\rho_x^{K-1}\rho_y^{L-1}}{(\sigma_x+\sigma_y-\theta)^{K+L-1}} h^{K+L-1} + o(h^{K+L+1}) \quad \text{as} \quad h \to \infty$$

Thus,

$$\begin{aligned}\mathbb{E}e^{\theta\frac{N'^{(h)}}{h}} &= \frac{(\sigma_x+\sigma_y)^{K+L-1}}{(\sigma_x+\sigma_y-\theta)^{K+L-1}} + o(1) \\ &\xrightarrow[h\to\infty]{} \frac{(\sigma_x+\sigma_y)^{K+L-1}}{(\sigma_x+\sigma_y-\theta)^{K+L-1}} = \mathbb{E}e^{\theta\hat{N}'}\end{aligned}$$

Thus by Lévy's Convergence Theorem the result holds (32). □

7.1 The 2 × 2 grid network

We now consider more explicitly the behaviour of 2×2 grid networks. The limiting distribution of a 2×2 grid network is characterized in the following proposition. This distribution differs from the distribution found in Theorem 1 and does not have the same heavy traffic limit as the queueing model. Interestingly we find that the limiting distribution can be expressed in terms of $K \times L$ independent exponential random variables conditioned on belonging to a certain linear subspace. This suggests a stronger form of collapse occurs for the proportionally fair stochastic flow level model. The following result is proved in Appendix A.2.

Proposition 4 *For each $h \in \mathbb{N}$, let $N^{(h)}$ have the stationary distribution of a proportionally fair stochastic flow level model on a 2×2 grid network with traffic intensities $\rho_x^{(h)} = \rho_x - \frac{\sigma_x}{h}$, $\rho_y^{(h)} = \rho_y - \frac{\sigma_y}{h}$ and $\rho_x + \rho_y = 1$. Then*

$$\frac{N^{(h)}}{h} \Rightarrow \tilde{N}$$

where

$$\tilde{N} = (\rho_x[\tilde{Q}_{(1,1)}+\tilde{Q}_{(1,2)}], \rho_x[\tilde{Q}_{(2,1)}+\tilde{Q}_{(2,2)}], \rho_y[\tilde{Q}_{(1,1)}+\tilde{Q}_{(2,1)}], \rho_y[\tilde{Q}_{(1,2)}+\tilde{Q}_{(2,2)}])$$

and $\tilde{Q}$ has the distribution of four independent exponential random variables constrained to the space $\{q \in \mathbb{R}_+^4 : q_{(1,1)} + q_{(2,2)} = q_{(2,1)} + q_{(1,2)}\}$, with density function

$$p(q) = C'\mathbb{I}[q_{(1,1)} + q_{(2,2)} = q_{(2,1)} + q_{(1,2)}]e^{-(\sigma_x+\sigma_y)(q_{(1,1)}+q_{(2,1)}+q_{(1,2)}+q_{(2,2)})}$$

where C' is a scaling constant and integration is taken over $q_{(1,1)}$, $q_{(1,2)}$ and $m := q_{(1,1)} + q_{(2,2)} = q_{(2,1)} + q_{(1,2)}$ with respect to the Lebesgue measure on $\mathbb{R}_+^3$.

Even though the stationary heavy traffic behaviour of the processor sharing queueing network and the proportionally fair flow model differ, the large deviations behaviour of both stationary distributions is the same.

Lemma 4 *If N has the stationary distribution of a proportionally fair stochastic flow level model on a 2×2 grid network with traffic intensities ρ_x, ρ_y and $\rho_x + \rho_y < 1$ then*

$$\begin{aligned}
&\lim_{h\to\infty} \frac{1}{h} \log \mathbb{P}(N = \lfloor hn \rfloor) \\
&= - \max_{\Lambda \in \mathbb{R}_+^4} \quad n_{x1} \log \frac{\Lambda_{x1}}{\rho_x} + n_{x2} \log \frac{\Lambda_{x2}}{\rho_x} + n_{y1} \log \frac{\Lambda_{y1}}{\rho_y} + n_{y2} \log \frac{\Lambda_{y2}}{\rho_y} \\
&\quad \text{subject to} \quad \Lambda_{xk} + \Lambda_{yl} \le 1 \quad \text{for} \quad k = 1,2, \quad l = 1,2.
\end{aligned}$$

Proof Using Stirling's approximation on the distribution of N gives

$$\begin{aligned}
&\lim_{h\to\infty} \frac{1}{h} \log \mathbb{P}(N = \lfloor hn \rfloor) \\
&= \lim_{h\to\infty} \frac{1}{h} \log \binom{\lfloor hn_{x1} \rfloor + \lfloor hn_{x2} \rfloor + \lfloor hn_{y1} \rfloor + \lfloor hn_{y2} \rfloor}{\lfloor hn_{x1} \rfloor + \lfloor hn_{x2} \rfloor} \rho_x^{\lfloor hn_{x1} \rfloor + \lfloor hn_{x2} \rfloor} \rho_y^{\lfloor hn_{y1} \rfloor + \lfloor hn_{y2} \rfloor} \\
&= \lim_{h\to\infty} \frac{1}{h} \big[h(n_{x1} + n_{x2} + n_{y1} + n_{y2}) \log(h(n_{x1} + n_{x2} + n_{y1} + n_{y2})) \\
&\qquad - h(n_{x1} + n_{x2}) \log \frac{h(n_{x1} + n_{x2})}{\rho_x} - h(n_{y1} + n_{y2}) \log \frac{h(n_{y1} + n_{y2})}{\rho_y} \big] \\
&= -(n_{x1} + n_{x2}) \log \frac{n_{x1} + n_{x2}}{\rho_x(n_{x1} + n_{x2} + n_{y1} + n_{y2})} \\
&\qquad - (n_{y1} + n_{y2}) \log \frac{n_{y1} + n_{y2}}{\rho_y(n_{x1} + n_{x2} + n_{y1} + n_{y2})}.
\end{aligned}$$

We know from Proposition 3 that this solves the proportional fair optimisation problem for a 2×2 grid network. □

8 Performance of modified proportional fairness

As we have seen for grid networks, product form networks of processor sharing queues do not capture the heavy traffic behaviour of proportional fairness. In order to study this behaviour in more detail we turn our attention to modified proportional fairness, an alternative allocation policy introduced in (22).

8.1 Modified proportional fairness

To define modified proportional fairness, we first let

$$\alpha(n) = \begin{cases} \sum_{s\in\mathcal{S}} n_s \log \Lambda_s^{PF}(n), & n \in \mathbb{Z}_+^S \\ \infty, & \text{otherwise.} \end{cases}$$

Also from expression (6.1) recall $\alpha_\rho(n)$. We define the modified proportionally fair allocation by

$$\Lambda_s^{MP}(n) = \exp\{\alpha(n) - \alpha(n - e_s)\}, \quad n \in \mathbb{R}_+^S. \tag{8.1}$$

In the following theorem we collect several results found in Massoulié (22) that relate modified proportional fairness and proportional fairness.

Theorem 5 *i) A stochastic flow level model operating under modified proportional fairness is reversible and has an invariant measure given by*

$$\pi_\rho^{MP}(n) = e^{-\alpha_\rho(n)}, \quad n \in \mathbb{Z}_+^S. \tag{8.2}$$

ii) The following large deviations relationship holds

$$\lim_{h\to\infty} \frac{1}{h} \log \pi_\rho^{MP}(\lfloor hn \rfloor) = -\alpha_\rho(n).$$

iii) For $s \in \mathcal{S}$ such that $n_s > 0$,

$$\Lambda_s^{MP}(hn) \xrightarrow[h\to\infty]{} \Lambda_s^{PF}(n).$$

Proof i) This can be verified with the detailed balance equations.
ii) This holds noting that $\alpha_\rho(\cdot)$ is continuous and that $\alpha_\rho(hn) = h\alpha_\rho(n)$.
iii) This can be verified using Lemmas 1 and 2 of Massoulié (22).

8.2 Modified proportional fairness in heavy traffic

We next explore how the invariant measure (8.2) behaves in heavy traffic. Let $\rho_s^{(h)} = \rho_s - \frac{\sigma_s}{h}$ for $s \in \mathcal{S}$ where as in Section 5 $\rho \in \mathbb{R}_+^S$ is Pareto efficient.

First note

$$\begin{aligned}\alpha_{\rho^{(h)}}(hn) - \alpha_\rho(hn) &= \alpha(hn) - \alpha(hn) - \sum_{s\in\mathcal{S}} hn_s \log \frac{\rho_s^{(h)}}{\rho_s} \\ &= -h \sum_{s\in\mathcal{S}} n_s \log(1 - \frac{\sigma_s}{h\rho_s}).\end{aligned}$$

Therefore

$$e^{-\alpha_{\rho^{(h)}}(hn)} = e^{-h\alpha_\rho(n)} \prod_{s\in\mathcal{S}} (1 - \frac{\sigma_s}{h\rho_s})^{hn_s}. \tag{8.3}$$

Now we consider the $e^{-h\alpha_\rho(n)}$ term in the above expression. We know that $\alpha_\rho(n) \geq 0$ since α_ρ is a rate function. Further,

$$\begin{aligned}\alpha_\rho(n) = 0 \iff & \sum_{s\in\mathcal{S}} n_s \log \frac{\Lambda_s^{PF}(n)}{\rho_s} = 0 \\ \iff & \sum_{s\in\mathcal{S}} n_s \log \Lambda_s^{PF}(n) = \sum_{s\in\mathcal{S}} n_s \log \rho_s \\ \iff & \Lambda_s^{PF}(n) = \rho_s,\end{aligned}$$

the final equivalence follows from the fact that ρ is feasible and achieves the optimal value of the objective function, and is therefore the unique optimum. Thus the only values of $n \in \mathbb{R}_+^S$ without a leading exponential decay term in expression (8.3) are those on the manifold $\mathcal{N}$ given by (5.4). From (8.3),

$$\begin{aligned}e^{-\alpha_{\rho^{(h)}}(hn)} &= e^{-h\alpha_\rho(n)} \prod_{s\in\mathcal{S}} (1 - \frac{\sigma_s}{h\rho_s})^{hn_s} \\ &\xrightarrow[h\to\infty]{} f(n) = \mathbb{I}[n \in \mathcal{N}] e^{-\sum_{s\in\mathcal{S}} n_s \frac{\sigma_s}{\rho_s}}.\end{aligned}$$

Note that the support of the density $f(n)$, the manifold $\mathcal{N}$, may well have dimension less than J. The density $f(n)$ is consistent with the results found for grid networks operating under proportional fairness in heavy traffic. Thus it seems plausible that the stationary distribution of proportional fairness in heavy traffic will agree with the above distribution. We present this as a conjecture in the following section.

9 Conjectures

The results from Sections 5 and 6 suggest a close relationship between the stationary behaviour of networks of processor sharing queues and proportionally fair stochastic flow level models, but the results in Section 7 put limitations on any such relationship. Based on these results we state a heavy traffic conjecture and a large deviations conjecture for the asymptotic behaviour of proportionally fair stochastic flow level models.

As in Theorem 1 we choose $\rho^{(h)}$ to be such that $\rho_s^{(h)} = \rho_s - \frac{\sigma_s}{h}$ for $s \in \mathcal{S}$ where $\rho \in \mathbb{R}_+^S$ is Pareto efficient. We also define $\mathcal{J}^* = \{j \in \mathcal{J} : \sum_{s\in\mathcal{S}} a_{js}\rho_s = C_j\}$. Define $A_{\mathcal{J}^*} = (a_{js} : j \in \mathcal{J}^*, s \in \mathcal{S})$, the submatrix of A formed by removing the j-th row for each $j \in \mathcal{J}\backslash\mathcal{J}^*$. Consider $Ker(A^t_{\mathcal{J}^*})$, the kernel of the transpose of $A_{\mathcal{J}^*}$ and let $\mathcal{K} = Ker(A^t_{\mathcal{J}^*})^\perp$ be its orthogonal complement. Note $\mathcal{K}$ is chosen so that the map $\mathcal{K} \to \mathcal{N}$, $q \mapsto (n : n_s = \sum_{j\in\mathcal{J}^*} q_j A_{js}\rho_s)$ is bijective.

Conjecture 1 *If, for each $h \in \mathbb{N}$, $N^{(h)}$ has the stationary distribution of a proportionally fair stochastic flow level model with traffic intensities $\rho^{(h)} = (\rho_s - \frac{\sigma_s}{h} : s \in \mathcal{S})$ then, as $h \to \infty$,*

$$\frac{N^{(h)}}{h} \Rightarrow \tilde{N},$$

where

$$\tilde{N}_s = \sum_{j\in\mathcal{J}^*} a_{js}\rho_s\tilde{Q}_j \qquad s \in \mathcal{S}.$$

Here, for each $j \in \mathcal{J}\backslash\mathcal{J}^*$, $\tilde{Q}_j = 0$, and $(\tilde{Q}_j : j \in \mathcal{J}^*)$ are independent exponential random variables with parameters $\sum_{s\in\mathcal{S}} a_{js}\sigma_s$ conditioned on belonging to the subspace $\mathcal{K}$; that is $(\tilde{Q}_j : j \in \mathcal{J}^*)$ has density

$$p(q) = C'\mathbb{I}[q \in \mathcal{K}]e^{-\sum_{j\in\mathcal{J}^*}\sum_{s\in\mathcal{S}} q_j a_{js}\sigma_s}$$

where C' is a scaling constant and integration is taken with respect to the Lebesgue measure on $\mathcal{K}$.

Note that if $A_{\mathcal{J}^*}$ is of full row rank then $\mathcal{K} = \mathbb{R}_+^{\mathcal{J}^*}$ and the conditioned random variables $(\tilde{Q}_j : j \in \mathcal{J}^*)$ of the conjecture remain independent exponentially distributed random variables.

It is interesting to compare this with a conjecture of Massoulié (22).

Conjecture 2 *If N has the stationary distribution of a proportionally fair stochastic flow level model with traffic intensities given by $\rho \in \mathbb{R}_+^S$ then, as $h \to \infty$, $\{\frac{N}{h}\}_{h\in\mathbb{N}}$ obeys a large deviation principle with good rate function*

$$\alpha_\rho(n) = \max_{\Lambda\in\mathbb{R}_+^S} \sum_{s\in\mathcal{S}} n_s \log\frac{\Lambda_s}{\rho_s} \quad \text{subject to} \quad \sum_{s\in\mathcal{S}} a_{js}\Lambda_s \le C_j,\ j \in \mathcal{J}. \tag{9.1}$$

That is for all Borel measurable $D \subset \mathbb{R}_+^S$ we have that

$$-\inf_{n\in D^\circ}\alpha_\rho(n) \le \liminf_{h\to\infty}\mathbb{P}(\frac{N}{h} \in D) \le \limsup_{h\to\infty}\mathbb{P}(\frac{N}{h} \in D) \le -\inf_{n\in\bar{D}}\alpha_\rho(n).$$

This conjecture suggests that the large deviations behaviour of the stationary distribution of a proportionally fair stochastic flow level model is unaffected by the row rank of the matrix A and that it agrees with the results found for networks of processor sharing queues in Section 6.

10 Concluding remarks

We conclude by recalling one of our aims, to provide insight into the performance consequences of resource pooling. A network of processor sharing queues has the remarkable property that the mean sojourn time in the network of an arriving document of class s is

$$\sum_{j\in\mathcal{J}} \frac{C_j}{C_j-\rho_j}\frac{a_{js}}{\mu_s}$$

where $\rho_j = \sum_{s\in\mathcal{S}} a_{js}\rho_s$ is the load on resource j, even when service requirements have arbitrary distributions (8; 14). The results we have reviewed are motivated by the possibility that an approximation of this form may be justified under proportionally fair multi-path routing, where $\mathcal{J}$ labels the set of pooled resources.

The heavy traffic results of (13) are suggestive, but the results concern the stationary distribution of a diffusion limit, rather than the limiting form of a stationary distribution, and also require the local traffic condition. The local traffic condition is, unfortunately, difficult to verify with multi-path routing. We have studied an example where the condition is not satisfied, the grid network of Section 7, and where the number of documents in the system has approximately an Erlang distribution, but arising from the sum of $K+L-1$, rather than KL, independent exponential random variables. Nevertheless, the behaviour of modified proportional fairness suggests a simple heavy traffic description of stationary proportionally fair flow models. This description coincides with that found in networks of processor sharing queues, under the less restrictive assumption that the matrix A is of full row rank.

A Appendix

A.1 Proof of Theorem 1

The characteristic function of $(M^{(h)}_{js} : j\in\mathcal{J}, s\in\mathcal{S})$ is

$$\mathbb{E}e^{i\sum_{j,s}\theta_{js}M^{(h)}_{js}} = \prod_{j\in\mathcal{J}}\left(\frac{C_j-\sum_{s\in\mathcal{S}} a_{js}\rho^{(h)}_s}{C_j-\sum_{s\in\mathcal{S}} a_{js}\rho^{(h)}_s e^{i\theta_{js}}}\right) \qquad \theta\in\mathbb{R}^{J\times S}_+.$$

The characteristic function of $(\hat{M}_j : j\in\mathcal{J})$ is

$$\mathbb{E}e^{i\sum_{j\in\mathcal{J}}\phi_j\hat{M}_j} = \prod_{j\in\mathcal{J}^*}\left(\frac{\sum_{s\in\mathcal{S}} a_{js}\sigma_s}{\sum_{s\in\mathcal{S}} a_{js}\sigma_s - iC_j\phi_j}\right) \qquad \phi\in\mathbb{R}^J_+.$$

Thus $(\frac{N_s^{(h)}}{h} : s \in \mathcal{S})$ has a characteristic function that converges in the following way

$$\begin{aligned}\mathbb{E}e^{i\sum_{s\in\mathcal{S}}\theta_s\frac{N_s^{(h)}}{h}} &= \prod_{j\in\mathcal{J}}\left(\frac{C_j-\sum_{s\in\mathcal{S}}a_{js}(\rho_s-\frac{\sigma_s}{h})}{C_j-\sum_{s\in\mathcal{S}}a_{js}(\rho_s-\frac{\sigma_s}{h})e^{i\frac{\theta_{js}}{h}}}\right)\\ &\xrightarrow[h\to\infty]{} \prod_{j\in\mathcal{J}^*}\left(\frac{\sum_{s\in\mathcal{S}}a_{js}\sigma_s}{\sum_{s\in\mathcal{S}}a_{js}\sigma_s - i\sum_{s\in\mathcal{S}}a_{js}\theta_s}\right)\\ &=\mathbb{E}e^{i\sum_{s\in\mathcal{S},j\in\mathcal{J}}a_{js}\theta_s\frac{\hat{M}_j}{C_j}} = \mathbb{E}e^{i\sum_{s\in\mathcal{S}}\theta_s\hat{N}_s}.\end{aligned}$$

Thus by Lévy's Convergence theorem (32) $\frac{N^{(h)}}{h} \Rightarrow \hat{N}$ as $h\to\infty$.

Now consider the proportionally fair optimisation problem

$$\max_{\Lambda\in\mathbb{R}_+^S}\sum_{s:\hat{N}_s>0}\hat{N}_s\log\Lambda_s \quad \text{subject to} \quad \sum_{s\in\mathcal{S}}a_{js}\Lambda_s\le C_j,\ \ j\in\mathcal{J}.$$

The proportional fairness optimality conditions (2.11-2.12) are satisfied by $\Lambda_s=\rho_s$ for $s\in\mathcal{S}$ and $q_j=\frac{\hat{M}_j}{C_j}$ for $j\in\mathcal{J}$. Therefore

$$\rho_s=\Lambda_s^{PF}(\hat{N}) \qquad s\in\mathcal{S}.$$

□

A.2 Proofs for grid networks

First we calculate the scaling constant for a $K\times L$ grid network.

Lemma 5

$$C(\rho)=\sum_{n\in\mathbb{Z}_+^{K+L}}\binom{\sum_{k=1}^K n_{xk}+\sum_{l=1}^L n_{yl}}{\sum_{k=1}^K n_{xk}}\rho_x^{\sum_{k=1}^K n_{xk}}\rho_y^{\sum_{l=1}^L n_{yl}} \tag{A.1}$$

$$=\sum_{k=1}^{K-1}\sum_{l=1}^{L-1}\binom{k+l}{k}\frac{\rho_x^k\rho_y^l}{(1-\rho_x-\rho_y)^{k+l+1}}. \tag{A.2}$$

Proof From (A.1) we can deduce the following.

$$\begin{aligned}&\sum_{n\in\mathbb{Z}_+^{K+L}}\binom{\sum_{k=1}^K n_{xk}+\sum_{l=1}^L n_{yl}}{\sum_{k=1}^K n_{xk}}\rho_x^{\sum_{k=1}^K n_{xk}}\rho_y^{\sum_{l=1}^L n_{yl}}\\ &=\sum_{n'=0}^{\infty}\sum_{\substack{(n_x,n_y):\\ n_x+n_y=n'}}\sum_{\substack{n\in\mathbb{Z}_+^{K+L}:\sum_k n_{xk}=n_x\\ \sum_l n_{yl}=n_y}}\binom{n_x+n_y}{n_x}\rho_x^{n_x}\rho_y^{n_y}\\ &=\sum_{n'=0}^{\infty}\sum_{\substack{(n_x,n_y):\\ n_x+n_y=n'}}\binom{n_x+K-1}{K-1}\binom{n_y+L-1}{L-1}\binom{n_x+n_y}{n_x}\rho_x^{n_x}\rho_y^{n_y}\\ &=\sum_{n'=0}^{\infty}\sum_{\substack{(n_x,n_y):\\ n_x+n_y=n'}}\left(\frac{\partial}{\partial\rho_x}\right)^{K-1}\left(\frac{\partial}{\partial\rho_y}\right)^{L-1}\left[\frac{\rho_x^{K-1}}{(K-1)!}\frac{\rho_y^{L-1}}{(L-1)!}\binom{n_x+n_y}{n_x}\rho_x^{n_x}\rho_y^{n_y}\right]\\ &=\left(\frac{\partial}{\partial\rho_x}\right)^{K-1}\left(\frac{\partial}{\partial\rho_y}\right)^{L-1}\left[\sum_{n'=0}^{\infty}\frac{\rho_x^{K-1}}{(K-1)!}\frac{\rho_y^{L-1}}{(L-1)!}(\rho_x+\rho_y)^{n'}\right]\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\partial}{\partial \rho_x}\right)^{K-1}\left(\frac{\partial}{\partial \rho_y}\right)^{L-1}\left[\frac{\rho_x^{K-1}}{(K-1)!}\frac{\rho_y^{L-1}}{(L-1)!}\frac{1}{(1-\rho_x-\rho_y)}\right] \\
&= \sum_{k=0}^{K-1}\sum_{l=0}^{L-1}\frac{1}{(K-1)!}\frac{\partial^{K-1-k}\rho_x^{K-1}}{\partial\rho_x^{K-1-k}}\cdot\frac{1}{(L-1)!}\frac{\partial^{L-1-l}\rho_y^{L-1}}{\partial\rho_y^{L-1-l}}\cdot\frac{\partial^{k+l}}{\partial\rho_x^k\partial\rho_y^l}\frac{1}{(1-\rho_x-\rho_y)} \\
&= \sum_{k=1}^{K-1}\sum_{l=1}^{L-1}\binom{k+l}{k}\frac{\rho_x^k\rho_y^l}{(1-\rho_x-\rho_y)^{k+l+1}}
\end{aligned}
$$

as required. □

To establish Proposition 4 we calculate the prelimit and limit characteristic functions and then prove convergence.

Lemma 6 *The characteristic function of the stationary distribution of a* 2×2 *grid network operating under proportional fairness with intensities* ρ_x*,* ρ_y *such that* $\rho_x+\rho_y<1$ *is*

$$
\mathbb{E}e^{i\theta_{x1}N_{x1}+i\theta_{x2}N_{x2}+i\theta_{y1}N_{y1}+i\theta_{y2}N_{y2}}
$$

$$
=\begin{cases}
\frac{C_\rho^{-1}}{(\rho_x e^{i\theta_{x1}}-\rho_x e^{i\theta_{x2}})(\rho_y e^{i\theta_{y1}}-\rho_y e^{i\theta_{y2}})}\times\Big[\frac{\rho_x e^{i\theta_{x1}}\rho_y e^{i\theta_{y1}}}{1-\rho_x e^{i\theta_{x1}}-\rho_y e^{i\theta_{y1}}}-\frac{\rho_x e^{i\theta_{x1}}\rho_y e^{i\theta_{y2}}}{1-\rho_x e^{i\theta_{x1}}-\rho_y e^{i\theta_{y2}}} \\
\qquad -\frac{\rho_x e^{i\theta_{x2}}\rho_y e^{i\theta_{y1}}}{1-\rho_x e^{i\theta_{x2}}-\rho_y e^{i\theta_{y1}}}+\frac{\rho_x e^{i\theta_{x2}}\rho_y e^{i\theta_{y2}}}{1-\rho_x e^{i\theta_{x2}}-\rho_y e^{i\theta_{y2}}}\Big] \\
C_\rho^{-1}\Big[\frac{2\rho_x e^{i\theta_{x2}}\rho_y e^{i\theta_{y2}}}{(1-\rho_x e^{i\theta_{x2}}-\rho_y e^{i\theta_{y2}})^3}+\frac{\rho_x e^{i\theta_{x2}}+\rho_y e^{i\theta_{y2}}}{(1-\rho_x e^{i\theta_{x2}}-\rho_y e^{i\theta_{y2}})^2}+\frac{1}{(1-\rho_x e^{i\theta_{x2}}-\rho_y e^{i\theta_{y2}})}\Big] \\
\frac{C_\rho^{-1}}{\rho_y(e^{i\theta_{y1}}-e^{i\theta_{y2}})}\Big[\frac{\rho_x e^{i\theta_{x1}}\rho_y e^{i\theta_{y1}}}{(1-\rho_x e^{i\theta_{x1}}-\rho_y e^{i\theta_{y1}})^2}-\frac{\rho_x e^{i\theta_{x1}}\rho_y e^{i\theta_{y2}}}{(1-\rho_x e^{i\theta_{x1}}-\rho_y e^{i\theta_{y2}})^2} \\
\qquad +\frac{\rho_x e^{i\theta_{y1}}}{(1-\rho_y e^{i\theta_{x1}}-\rho_x e^{i\theta_{y1}})}-\frac{\rho_x e^{i\theta_{y2}}}{(1-\rho_y e^{i\theta_{x1}}-\rho_x e^{i\theta_{y2}})}\Big] \\
\frac{C_\rho^{-1}}{\rho_x(e^{i\theta_{x1}}-e^{i\theta_{x2}})}\Big[\frac{\rho_y e^{i\theta_{y1}}\rho_x e^{i\theta_{x1}}}{(1-\rho_y e^{i\theta_{y1}}-\rho_x e^{i\theta_{x1}})^2}-\frac{\rho_y e^{i\theta_{y1}}\rho_x e^{i\theta_{x2}}}{(1-\rho_y e^{i\theta_{y1}}-\rho_x e^{i\theta_{x2}})^2} \\
\qquad +\frac{\rho_x e^{i\theta_{x1}}}{(1-\rho_y e^{i\theta_{y1}}-\rho_x e^{i\theta_{x1}})}-\frac{\rho_x e^{i\theta_{x2}}}{(1-\rho_y e^{i\theta_{y1}}-\rho_x e^{i\theta_{x2}})}\Big]
\end{cases}
$$

where the 4 cases above correspond respectively to the cases: $\theta_{x1}\neq\theta_{x2}$ *and* $\theta_{y1}\neq\theta_{y2}$*;* $\theta_{x1}=\theta_{x2}$ *and* $\theta_{y1}=\theta_{y2}$*;* $\theta_{x1}=\theta_{x2}$ *and* $\theta_{y1}\neq\theta_{y2}$*, and* $\theta_{x1}\neq\theta_{x2}$ *and* $\theta_{y1}=\theta_{y2}$*.*

Proof Suppose $\theta_{x1}\neq\theta_{x2}$ and $\theta_{y1}\neq\theta_{y2}$.

$$
\begin{aligned}
&C_\rho\mathbb{E}e^{i\theta_{x1}N_{x1}+i\theta_{x2}N_{x2}+i\theta_{y1}N_{y1}+i\theta_{y2}N_{y2}} \\
&=\sum_{n\geq 0}\binom{n_{x1}+n_{x2}+n_{y1}+n_{y2}}{n_{x1}+n_{x2}}(\rho_x e^{i\theta_{x1}})^{n_{x1}}(\rho_x e^{i\theta_{x2}})^{n_{x2}}(\rho_y e^{i\theta_{y1}})^{n_{y1}}(\rho_y e^{i\theta_{y2}})^{n_{y2}} \\
&=\sum_{K=0}^{\infty}\sum_{k=0}^{K}\binom{K}{k}\sum_{\substack{n:n_{x1}+n_{x2}=k\\ n_{y1}+n_{y2}=K-k}}(\rho_x e^{i\theta_{x2}})^k(\rho_y e^{i\theta_{y2}})^{K-k}(e^{i\theta_{x1}-i\theta_{x2}})^{n_1}(e^{i\theta_{y1}+i\theta_{y2}})^{n_{y1}} \\
&=\sum_{K=0}^{\infty}\sum_{k=0}^{K}\binom{K}{k}(\rho_x e^{i\theta_{x2}})^k(\rho_y e^{i\theta_{y2}})^{K-k}\left(\frac{1-e^{(i\theta_{x1}-i\theta_{x2})(k+1)}}{1-e^{i\theta_{x1}-i\theta_{x2}}}\right)\left(\frac{1-e^{(i\theta_{y1}-i\theta_{y2})(K-k+1)}}{1-e^{i\theta_{y1}-i\theta_{y2}}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{1}{e^{i\theta_{x2}} - e^{i\theta_{x1}}}\right)\left(\frac{1}{e^{i\theta_{y2}} - e^{i\theta_{y1}}}\right) \\
&\quad \times \sum_{K=0}^{\infty}\sum_{k=0}^{K}\binom{K}{k}\Big[e^{i\theta_{x2}}e^{i\theta_{y2}}(\rho_x e^{i\theta_{x2}})^k(\rho_y e^{i\theta_{y2}})^{K-k} - e^{i\theta_{x1}}e^{i\theta_{y2}}(\rho_x e^{i\theta_{x1}})^k(\rho_y e^{i\theta_{y2}})^{K-k} \\
&\qquad\qquad - e^{i\theta_{x2}}e^{i\theta_{y1}}(\rho_x e^{i\theta_{x2}})^k(\rho_y e^{i\theta_{y1}})^{K-k} + e^{i\theta_{x1}}e^{i\theta_{y1}}(\rho_x e^{i\theta_{x1}})^k(\rho_y e^{i\theta_{y1}})^{K-k}\Big] \\
&= \left(\frac{1}{e^{i\theta_{x2}} - e^{i\theta_{x1}}}\right)\left(\frac{1}{e^{i\theta_{y2}} - e^{i\theta_{y3}}}\right) \\
&\quad \times \sum_{K=0}^{\infty}\Big[e^{i\theta_{x2}}e^{i\theta_{y2}}(\rho_x e^{i\theta_{x2}} + \rho_y e^{i\theta_{y2}})^K - e^{i\theta_{x1}}e^{i\theta_{y2}}(\rho_x e^{i\theta_{x1}} + \rho_y e^{i\theta_{y2}})^K \\
&\qquad\qquad - e^{i\theta_{x2}}e^{i\theta_{y1}}(\rho_x e^{i\theta_{x2}} + \rho_y e^{i\theta_{y1}})^K + e^{i\theta_{x1}}e^{i\theta_{y1}}(\rho_x e^{i\theta_{x1}} + \rho_y e^{i\theta_{y1}})^K\Big] \\
&= \left(\frac{1}{e^{i\theta_{x2}} - e^{i\theta_{x1}}}\right)\left(\frac{1}{e^{i\theta_{y2}} - e^{i\theta_{y1}}}\right)\Big[\frac{e^{i\theta_{x2}}e^{i\theta_{y2}}}{1-\rho_x e^{i\theta_{x2}} - \rho_y e^{i\theta_{y2}}} - \frac{e^{i\theta_{x1}}e^{i\theta_{y2}}}{1-\rho_x e^{i\theta_{x1}} - \rho_y e^{i\theta_{y2}}} \\
&\qquad\qquad - \frac{e^{i\theta_{x2}}e^{i\theta_{y1}}}{1-\rho_x e^{i\theta_{x2}} - \rho_y e^{i\theta_{y1}}} + \frac{e^{i\theta_{x1}}e^{i\theta_{y1}}}{1-\rho_x e^{i\theta_{x1}} - \rho_y e^{i\theta_{y1}}}\Big].
\end{aligned}
$$

This gives the $\theta_{x1} \neq \theta_{x2}$ and $\theta_{y1} \neq \theta_{y2}$ case. The $\theta_{x1} = \theta_{x2}$ and $\theta_{y1} = \theta_{y2}$ case is given by Lemma 5. The remaining cases have a proof that is very similar to these two cases. □

Lemma 7 *Let $\tilde{N}$ be a random variable in $\mathbb{R}^4_+$ with a density given by exponential random variables constrained to a linear subspace*

$$
p(n) = \tilde{C}_\rho^{-1}\mathbb{I}\left[\frac{n_{x1}+n_{x2}}{\rho_x} = \frac{n_{y1}+n_{y2}}{\rho_y}\right]e^{-\frac{\sigma_x}{\rho_x}(n_{x1}+n_{x2})}e^{-\frac{\sigma_y}{\rho_y}(n_{y1}+n_{y2})}
$$

where $\tilde{C}_\rho = \frac{2\rho_x\rho_y}{(\sigma_x+\sigma_y)^3}$ and integration is assumed to be taken over $dn_{x1}dn_{y1}dm$ with $m = \frac{n_{x1}+n_{x2}}{\rho_x} = \frac{n_{y1}+n_{y2}}{\rho_y}$. Then $\tilde{N}$ has characteristic function

$$
\mathbb{E}e^{i\theta\cdot\tilde{N}} = \begin{cases}
\frac{\tilde{C}_\rho^{1}}{(i\theta_{x1}-i\theta_{x2})(i\theta_{y1}-i\theta_{y2})}\Big[\frac{1}{\sigma_x+\sigma_y-\rho_x i\theta_{x2}-\rho_y i\theta_{y2}} - \frac{1}{\sigma_x+\sigma_y-\rho_x i\theta_{x1}-\rho_y i\theta_{y2}} \\
\qquad -\frac{1}{\sigma_x+\sigma_y-\rho_x i\theta_{x2}-\rho_y i\theta_{y1}} + \frac{1}{\sigma_x+\sigma_y-\rho_x i\theta_{x1}-\rho_y i\theta_{3}y1}\Big] \\
\frac{2\tilde{C}_\rho^{-1}\rho_x\rho_y}{(\sigma_x+\sigma_y-\rho_x i\theta_{x1}-\rho_y i\theta_{y1})^3} \\
\frac{\tilde{C}_\rho^{-1}}{(i\theta_{y1}-i\theta_{y2})}\left[\frac{\rho_x}{(\sigma_x+\sigma_y-\rho_x i\theta_{x1}-\rho_y i\theta_{y1})^2} - \frac{\rho_x}{(\sigma_x+\sigma_y-\rho_x i\theta_{x1}-\rho_y i\theta_{y2})^2}\right] \\
\frac{\tilde{C}_\rho^{-1}}{(i\theta_{x1}-i\theta_{x2})}\left[\frac{\rho_y}{(\sigma_x+\sigma_y-\rho_x i\theta_{x1}-\rho_y i\theta_{y1})^2} - \frac{\rho_y}{(\sigma_x+\sigma_y-\rho_x i\theta_{x2}-\rho_y i\theta_{y1})^2}\right]
\end{cases}
$$

where the four cases above correspond respectively to the cases: $\theta_{x1} \neq \theta_{x2}$ and $\theta_{y1} \neq \theta_{y2}$; $\theta_{x1} = \theta_{x2}$ and $\theta_{y1} = \theta_{y2}$; $\theta_{x1} = \theta_{x2}$ and $\theta_{y1} \neq \theta_{y2}$, and $\theta_{x1} \neq \theta_{x2}$ and $\theta_{y1} = \theta_{y2}$.

Proof Once again we only consider $\theta_{x1} \neq \theta_{x2}$, $\theta_{y1} \neq \theta_{y2}$ case (the other cases follow similarly).

$$
\begin{aligned}
&\mathbb{E}e^{i\theta\cdot\tilde{N}} \\
&= \tilde{C}_\rho^{-1}\int_0^\infty\int_0^{m\rho_y}\int_0^{m\rho_x} e^{i\theta_{x1}n_{x1}}e^{i\theta_{x2}(\rho_x m-n_{x1})}e^{i\theta_{y1}n_{y1}}e^{i\theta_{y2}(\rho_y m-n_{y1})}e^{-(\sigma_x+\sigma_y)m}dn_{x1}dn_{y1}dm \\
&= \tilde{C}_\rho^{-1}\int_0^\infty e^{(i\theta_{x2}\rho_x-\sigma_x)m}e^{(i\theta_{y2}\rho_y-\sigma_y)m}\int_0^{m\rho_x}\int_0^{m\rho_y} e^{-(i\theta_{x2}-i\theta_{x1})n_1}e^{-(i\theta_{y2}-i\theta_{x1})n_{y1}}dn_{x1}dn_{y1}dm \\
&= \tilde{C}_\rho^{-1}\int_0^\infty e^{(i\theta_{x2}\rho_x-\sigma_x)m}e^{(i\theta_{y2}\rho_y-\sigma_y)m}\frac{1-e^{-(i\theta_{x2}-i\theta_{x1})\rho_x m}}{i\theta_{x2}-i\theta_{x1}}\frac{1-e^{-(i\theta_{y2}-i\theta_{y1})\rho_y m}}{i\theta_{y2}-i\theta_{y1}}dm \\
&= \frac{\tilde{C}_\rho^{-1}}{(i\theta_{x2}-i\theta_{x1})(i\theta_{y2}-i\theta_{y1})}\int_0^\infty e^{(i\theta_{x2}\rho_x+i\theta_{y2}\rho_y-\sigma_x-\sigma_y)m}-e^{(i\theta_{x1}\rho_x+i\theta_{y2}\rho_y-\sigma_x-\sigma_y)m} \\
&\qquad -e^{(i\theta_{x2}\rho_x+i\theta_{y1}\rho_y-\sigma_x-\sigma_y)m}+e^{(i\theta_{x1}\rho_x+i\theta_{y1}\rho_y-\sigma_x-\sigma_y)m}dm \\
&= \frac{\tilde{C}_\rho^{-1}}{(i\theta_{x2}-i\theta_{x1})(i\theta_{y2}-i\theta_{y1})}\Bigg[\frac{1}{\sigma_x+\sigma_y-i\theta_{x2}\rho_x-i\theta_{y2}\rho_y}-\frac{1}{\sigma_x+\sigma_y-i\theta_{x1}\rho_x-i\theta_{y2}\rho_y} \\
&\qquad -\frac{1}{\sigma_x+\sigma_y-i\theta_{x2}\rho_x-i\theta_{y1}\rho_y}+\frac{1}{\sigma_x+\sigma_y-i\theta_{x1}\rho_x-i\theta_{y1}\rho_y}\Bigg]
\end{aligned}
$$

as required. □

Proof (of Proposition 4) We consider the limit behaviour of the characteristic function of $\frac{N^h}{h}$. We only consider the case where $\theta_{x1} \neq \theta_{x2}$ and $\theta_{y1} \neq \theta_{y2}$. The other cases follow similarly. For $\theta \in \mathbb{R}^4$ with $\theta_{x1} \neq \theta_{x2}$ and $\theta_{y1} \neq \theta_{y2}$,

$$
\begin{aligned}
&C_\rho^{-1}\mathbb{E}e^{i\theta_{x1}\frac{N^{(h)}x1}{h}+i\theta_{x2}\frac{N^{(h)}_{x2}}{h}+i\theta_{y1}\frac{N^{(h)}y1}{h}+i\theta_{y2}\frac{N^{(h)}_{y2}}{h}} \\
&= \left(\frac{1}{e^{\frac{i\theta_{x1}}{h}}-e^{\frac{i\theta_{x2}}{h}}}\right)\left(\frac{1}{e^{\frac{i\theta_{y1}}{h}}-e^{\frac{i\theta_{y2}}{h}}}\right)\times\Bigg[\frac{e^{\frac{i\theta_{x2}}{h}}e^{\frac{i\theta_{y2}}{h}}}{1-\rho_x^{(h)}e^{\frac{i\theta_{x2}}{h}}-\rho_y^{(h)}e^{\frac{i\theta_{y2}}{h}}}-\frac{e^{\frac{i\theta_{x1}}{h}}e^{\frac{i\theta_{y2}}{h}}}{1-\rho_x^{(h)}e^{\frac{i\theta_{x1}}{h}}-\rho_y^{(h)}e^{\frac{i\theta_{y2}}{h}}} \\
&\qquad -\frac{e^{\frac{i\theta_{x2}}{h}}e^{\frac{i\theta_{y1}}{h}}}{1-\rho_x^{(h)}e^{\frac{i\theta_{x2}}{h}}-\rho_y^{(h)}e^{\frac{i\theta_{y1}}{h}}}+\frac{e^{\frac{i\theta_{x1}}{h}}e^{\frac{i\theta_{y1}}{h}}}{1-\rho_x^{(h)}e^{\frac{i\theta_{x1}}{h}}-\rho_y^{(h)}e^{\frac{i\theta_{y1}}{h}}}\Bigg] \\
&= \frac{h^3}{(i\theta_{x1}-i\theta_{x2})(i\theta_{y1}-i\theta_{y2})}\Bigg[\frac{1}{\sigma_x+\sigma_y-i\theta_{x2}\rho_x-i\theta_{y2}\rho_y}-\frac{1}{\sigma_x+\sigma_y-i\theta_{x1}\rho_x-i\theta_{y2}\rho_y} \\
&\qquad -\frac{1}{\sigma_x+\sigma_y-i\theta_{x2}\rho_x-i\theta_{y1}\rho_y}+\frac{1}{\sigma_x+\sigma_y-i\theta_{x1}\rho_x-i\theta_{y1}\rho_y}\Bigg]+o(h^3).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
C_{\rho^{(h)}} &= \frac{2\rho_x^{(h)}\rho_y^{(h)}}{(1-\rho_x^{(h)}-\rho_y^{(h)})^3}+\frac{\rho_x^{(h)}+\rho_y^{(h)}}{(1-\rho_x^{(h)}-\rho_y^{(h)})^2}+\frac{1}{1-\rho_x^{(h)}-\rho_y^{(h)}} \\
&= \frac{2\rho_x\rho_y}{(\sigma_x+\sigma_y)^3}h^3+o(h^3).
\end{aligned}
$$

Thus comparing these two expressions gives that

$$
\mathbb{E}e^{i\theta\cdot\frac{N^{(h)}}{h}} \xrightarrow[h\to\infty]{} \mathbb{E}e^{i\theta\cdot\tilde{N}}
$$

for all $\theta_{x1} \neq \theta_{x2}$ and $\theta_{y1} \neq \theta_{y2}$. The other cases follow similarly. Thus, by Lévy's Convergence Theorem (32), $\frac{N^{(h)}}{h} \Rightarrow \tilde{N}$ where $\tilde{N}$ has density as given in Lemma 7.

The linear map given by matrix

$$\begin{pmatrix} n_{x1} \\ n_{x2} \\ n_{y1} \\ n_{y2} \end{pmatrix} = \begin{pmatrix} \rho_x & \rho_x & 0 & 0 \\ 0 & 0 & \rho_x & \rho_x \\ \rho_y & 0 & \rho_y & 0 \\ 0 & \rho_y & 0 & \rho_y \end{pmatrix} \begin{pmatrix} q_{(1,1)} \\ q_{(1,2)} \\ q_{(2,1)} \\ q_{(2,2)} \end{pmatrix}$$

is bijective from $\{q \in \mathbb{R}^4_+ : q_{(1,1)} + q_{(2,2)} = q_{(1,2)} + q_{(2,2)}\}$ to $\{n \in \mathbb{R}^4_+ : \frac{n_{x1}+n_{x2}}{\rho_x} = \frac{n_{y1}+n_{y2}}{\rho_y}\}$. Therefore we can express the density function from Lemma 7 in the form given in Proposition 4. □

References

1. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
2. Bonald, T., Massoulié, L.: Impact of fairness on Internet performance. Proc. of ACM Sigmetrics **29**, 82–91 (2001)
3. Bonald, T., Massoulié, L., Proutière, A., Virtamo, J.: A queueing analysis of max-min fairness, proportional fairness and balanced fairness. Queueing Systems, **53**, 65-84 (2006)
4. Bonald, T., Proutière, A.: Insensitive Bandwidth Sharing in Data Networks. Queueing Systems **44**, 69–100 (2003)
5. Bonald, T., Proutière, A.: On performance bounds for balanced fairness. Performance Evaluation **55**, 25–50 (2004)
6. Bramson, M.: State space collapse with application to heavy traffic limits for multiclass queueing networks. Queueing Systems **30**, 89–148 (1998)
7. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization. Springer, New York (2001)
8. Cohen, J.W.: The multiple phase service network with generalized processor sharing. Acta Informatica **12**, 245–284 (1979)
9. Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications. Springer, New York (1998)
10. De Veciana, G., Lee, T.J. and Konstantopoulos, T.: Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Trans. Networking* **9**, 2–14. (2001)
11. Dukkipati, N., Kobayashi, M., Zhang-Shen, R., McKeown, N.: Processor sharing flows in the Internet. Thirteenth International Workshop on Quality of Service (IWQoS), Passau, Germany (2005)
12. Han, H., Shakkottai, S., Hollot, C.V., Srikant, R., Towsley, D.: Multi-path TCP: a joint congestion control and routing scheme to exploit path diversity on the Internet. IEEE/ACM Trans. Networking **14**, 1260-1271 (2006)
13. Kang, W.N., Kelly, F.P., Lee, N.H., Williams, R. J.: State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. To appear in Ann. Appl. Probab. (2009)
14. Kelly, F.P.: Reversibility and Stochastic Networks. Wiley, Chicester (1979)
15. Kelly, F.P.: On a class of approximations for closed queueing networks. Queueing Systems **4**, 69–76 (1989)
16. Kelly, F.P.: Charging and rate control for elastic traffic. European Transactions on Telecommunications **8**, 33–37 (1997)
17. Kelly, F.P., Laws, C.N.: Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. Queueing Systems **13**, 47-86 (1993)
18. Kelly, F.P., Williams, R.J.: Fluid model for a network operating under a fair bandwidth-sharing policy. Ann. Appl. Probab. **14**, 1055–1083 (2004)
19. Key, P., Massoulié, L.: Fluid models of integrated traffic and multipath routing. Queueing Systems **53**, 85–98 (2006)
20. Kleinrock, L.: Time-shared systems: a theoretical treatment. *J.A.C.M.* **14**, 242–261 (1967)
21. Le Boudec, J.-Y., Radunovic, B.: Rate performance objectives of multihop wireless networks. IEEE Trans. Mob. Comput. **3**, 334-349, (2004)

22. Massoulié, L.: Structural properties of proportional fairness: stability and insensitivity. Ann. Appl. Probab. **17**, 809–839 (2007)
23. Massoulié, L., Roberts, J.: Bandwidth sharing and admission control for elastic traffic. Telecommunication Systems **15**, 185–201 (1998)
24. Massoulié, L., Roberts, J.: Bandwidth sharing: objectives and algorithms. IEEE Infocom **10**(3), 320–328 (1999)
25. Mo, J., Walrand, J.: Fair end-to-end window-based congestion control. IEEE/ACM Trans. Networking **8**, 556-567 (2000)
26. Pittel, B.: Closed exponential networks of queues with saturation. Math. Oper. Res. **4**, 357–378 (1979)
27. Proutière, A.: Insensitivity and stochastic bounds in queueing networks-Application to flow level traffic modelling in telecommunication networks. PhD thesis, Ecole Doctorale de l'Ecole Polytechnique (2003)
28. Schweitzer, P.J.: Approximate analysis of multiclass closed networks of queues. Proceedings of the international conference on stochastic control and optimization, Free University, Amsterdam, 1–5 (1979)
29. Srikant, R.: The Mathematics of Internet Congestion Control. Birkhauser, Boston (2004)
30. Walton, N.S.: Proportional fairness and its relationship with multi-class queueing networks. To appear in Ann. Appl. Probab. (2009)
31. Walton N.S.: Utility Optimisation Arising in Congested Queueing Networks. PhD thesis, University of Cambridge (2009)
32. Williams, D.: Probability with Martingales. Cambridge University Press, Cambridge (1991)
33. Wischik, D., Handley, M., Bagnulo Braun, M.: The resource pooling principle. Comput. Commun. Rev. **38**(5), 47-52 (2008)