



Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Fluid Limits for Bandwidth-Sharing Networks with Rate Constraints

Maria Remerova, Josh Reed, Bert Zwart

To cite this article:

Maria Remerova, Josh Reed, Bert Zwart (2014) Fluid Limits for Bandwidth-Sharing Networks with Rate Constraints. *Mathematics of Operations Research* 39(3):746-774. <http://dx.doi.org/10.1287/moor.2013.0641>

Full terms and conditions of use: <http://pubsonline.informs.org/page/terms-and-conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2014, INFORMS

Please scroll down for article—it is on subsequent pages



INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Fluid Limits for Bandwidth-Sharing Networks with Rate Constraints

Maria Remerova

Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, maria@remerova.ru

Josh Reed

Stern School of Business, New York University New York, New York 10012, jreed@stern.nyu.edu

Bert Zwart

Centrum Wiskunde & Informatica; EURANDOM, Eindhoven, The Netherlands; Vrije Universiteit Amsterdam, The Netherlands; and Georgia Institute of Technology, Atlanta, Georgia 30332, bert.zwart@cwi.nl

Bandwidth-sharing networks as introduced by Roberts and Massoulié [Roberts JW, Massoulié L (1998) Bandwidth sharing and admission control for elastic traffic. *Proc. ITC Specialist Seminar, Yokohama, Japan*], Massoulié and Roberts [Massoulié L, Roberts JW (1999) Bandwidth sharing: Objectives and algorithms. *Proc. IEEE Infocom*. (Books in Statistics, New York), 1395–1403] model the dynamic interaction among an evolving population of elastic flows competing for several links. With policies based on optimization procedures, such models are of interest both from a queueing theory and operations research perspective. In the present paper, we focus on bandwidth-sharing networks with capacities and arrival rates of a large order of magnitude compared to transfer rates of individual flows. This regime is standard in practice. In particular, we extend previous work by Reed and Zwart [Reed J, Zwart B (2010) Limit theorems for bandwidth-sharing networks with rate constraints. Revised, preprint <http://people.stern.nyu.edu/jreed/Papers/BARvised.pdf>] on fluid approximations for such networks: we allow interarrival times, flow sizes, and patient times (i.e., abandonment times measured from the arrival epochs) to be generally distributed, rather than exponentially distributed. We also develop polynomial-time computable fixed-point approximations for stationary distributions of bandwidth-sharing networks, and suggest new techniques for deriving these types of results.

Keywords: bandwidth sharing; rate constraints; patience times; large capacity scaling; fluid limits; fixed-point approximations

MSC2000 subject classification: Primary: 60K25, 60K30, 60F17, 60G57; secondary: 90B15, 90B22

OR/MS subject classification: Primary: queues, networks, limit theorems, approximations; secondary: probability, stochastic model applications

History: Received March 21, 2012; revised December 18, 2013. Published online in *Articles in Advance* February 21, 2014.

1. Introduction. Bandwidth-sharing policies as introduced by Roberts and Massoulié [27] and Massoulié and Roberts [23] dynamically distribute network resources among a changing population of users. Processor sharing is an example of such a policy and assumes a single resource. Bandwidth-sharing networks are of great research and practical interest. Along with the basic application in telecommunications media, e.g., Internet congestion control, they also have recently been suggested as a tool in analyzing problems in road traffic (Kelly and Williams [21]).

The main issues in bandwidth-sharing related research are stability conditions and performance evaluation. A variety of results regarding the first topic may be found in de Veciana et al. [9, 10], Bonald and Massoulié [4], Mo and Walrand [24], Massoulié [22], Bramson [7], Gromoll and Williams [13], and Chiang et al. [8]. As for the second topic, for special combinations of network topologies and bandwidth-sharing policies, the network stationary distribution may be shown to be of a product form insensitive to the flow size distribution; see Bonald et al. [5]. However, in general, approximation methods must be used, which is the subject matter of the present paper. Fundamental papers on fluid limit approximations for bandwidth-sharing networks are Kelly and Williams [20] and Gromoll and Williams [14], some more results on fluid and diffusion approximations are to be found in Borst et al. [6], Egorova et al. [11], Kang et al. [19], and Ye and Yao [29, 30]. The latter works ignore the fact that generally in practice the maximum service rate of an individual user is constrained, as has been pointed out by Roberts [26].

To the best of our knowledge, Ayesta and Mandjes [2] were the first to deal with fluid and diffusion approximations of bandwidth-sharing networks with rate limitations. They consider two specific settings first without rate constraints, and then they truncate the capacity constraints at the rate maxima. Reed and Zwart [25] develop a different approach in the context of general bandwidth-sharing networks. They incorporate the rate constraints into the network utility maximization procedure that defines bandwidth allocations. Thus, users operating below the maximal rate are allowed to take up the bandwidth that is not used by other rate constrained users, and bandwidth allocations are Pareto optimal. Another interesting feature of this work is the scaling regime. In contrast to the papers mentioned above, which mostly focus on the large-time properties of networks with fixed-order parameters, Reed and Zwart [25] view networks on a large-time scale letting arrival rates and capacities

grow large. This large capacity scaling reflects the fact that overall network capacity and individual user rate constraints may be of different orders of magnitude. For example, it is common that Internet providers set download speed limitations for individual users that are typically measured in megabits per second, whereas network capacities are measured in gigabits or terabits per second.

The framework of Reed and Zwart [25] is rather comprehensive. In particular, it allows abandonments of flows: each flow knows how long it can stay in the system and abandons as soon as its service is finished or its patience time expires, whichever happens earlier. The present paper builds upon Reed and Zwart [25] by relaxing its stochastic assumptions: we assume general distribution for interarrival times and general joint distribution for the size and patience time of a flow (in particular, the flow size and patience time are allowed to be dependent), whereas Reed and Zwart [25] assume a Markovian setting with independent arrivals, flow sizes, and patience times. We study the behavior of bandwidth-sharing networks in terms of measure-valued processes that are called state descriptors and that keep track of residual flow sizes and residual patience times. The first main result of the paper is a fluid limit theorem (it generalizes the fluid limit result of Reed and Zwart [25] to non-Markovian stochastic assumptions). We propose a fluid model, or a formal deterministic approximation of the stochastic bandwidth-sharing model, and show that the scaled state descriptors are tight with all weak limit points a.s. solving the fluid model equation. We provide a sufficient condition for the fluid model to have a unique solution, which converts tightness of the scaled state descriptors into convergence to this fluid model solution (FMS). In the sense of techniques used in the proofs, this part of the paper is closely related to previous work on bandwidth sharing (Gromoll and Williams [14]), processor sharing with impatience (Gromoll et al. [15]), and bandwidth sharing in overload (Borst et al. [6], Egorova et al. [11]). The rate constraints play a crucial role in adopting these techniques. For example, the proof of convergence to fluid model solutions in Gromoll et al. [15] requires an additional assumption of overload to eliminate problems at zero. However, in our case, because of the rate constraints, the network never empties, and the load conditions become irrelevant.

Our second main result, which is a new type of result for bandwidth-sharing networks, is convergence of the scaled network stationary distribution to the fixed point of the fluid model, provided the fixed point is unique. There is a similar result by Kang and Ramanan [18] for a call center model, but the techniques of Kang and Ramanan [18] are different than ours. Applying the approach of Borst et al. [6], we prove that in many cases the fixed point can be found by solving an optimization problem with a strictly concave objective function and a polyhedral constraint set, and thus is unique and computable in polynomial time. We also construct an example with multiple fixed points, which is a feature that is distinctive from earlier cited works. Besides proving new results for the particular model of bandwidth sharing, we also suggest new ideas and believe that they can be adjusted to other models, too. In particular, we derive equations for asymptotic bounds for fluid model solutions (see Theorem 3) that can be solved for a wide class of networks, and then asymptotic stability of the fixed point can be shown. Another interesting idea is that, in the stationary regime, the properties of a network depend on newly arriving flows only, since all initial flows are gone after some point (see Lemma 2). Throughout this part of the paper, we assume Poisson arrivals, since that guarantees existence of a unique stationary distribution. Poisson arrivals also imply $M/G/\infty$ bounds that are exploited heavily in the proofs.

The structure of the paper is as follows. Section 2 describes the stochastic bandwidth-sharing model, and §3 introduces its deterministic analogue, the fluid model. Also §3 states sufficient conditions for a fluid model solution to be unique, and for a fixed fluid model solution to be unique and asymptotically stable. Sections 4 and 5 discuss convergence of the scaled state descriptor and its stationary distribution to the fluid model and its fixed point, respectively. Sections 6, 7, and 8 contain the proofs of the statements from §§3, 4, and 5. The appendix proves auxiliary results. In the remainder of this section, we list the notation we use throughout the paper.

Notation. To introduce the notation, we use the signs $=:$ and $:=$.

The standard sets are denoted as follows: the reals $\mathbb{R} = (-\infty, \infty)$, the nonnegative reals $\mathbb{R}_+ = [0, \infty)$, the positive reals $(0, \infty)$, the nonnegative integers $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, and the natural numbers $\mathbb{N} = \{1, 2, \dots\}$.

The signs \wedge and \vee stand for minimum and maximum, respectively. For $x \in \mathbb{R}$, $x^+ := x \vee 0$.

The signs $\underline{\lim}$ and $\overline{\lim}$ denote the lower and upper limits of a sequence of numbers.

The coordinates of a vector from a set S^I are denoted by the same symbol as the vector with lower indices $1, \dots, I$ added. If a vector has a superscript, tilde sign, or overlining, they remain in its coordinates. For example $\bar{x}^0 \in S^I$, $\bar{x}^0 = (\bar{x}_1^0, \dots, \bar{x}_I^0)$. The space \mathbb{R}^I is endowed with the supremum norm $\|x\| := \max_{1 \leq i \leq I} |x_i|$. Vector inequalities hold coordinate-wise. The coordinate-wise product of vectors of the same dimensionality I is $x * y := (x_1 y_1, \dots, x_I y_I)$.

The signs \Rightarrow , $\stackrel{d}{=}$, and \leq_{st} stand for convergence in distribution, equality in distribution, and stochastic order, respectively. Recall that, for random variables (r.v.'s) X and X' , $X \leq_{st} X'$ if $\mathbb{P}\{X > x\} \leq \mathbb{P}\{X' > x\}$ for all $x \in \mathbb{R}$. The notation $\Pi(\lambda)$, $\lambda \in (0, \infty)$, stands for the Poisson distribution with parameter λ .

For metric spaces S and S' , denote by $\mathbf{C}_{S \rightarrow S'}$ the space of continuous functions $f: S \rightarrow S'$. By $\mathbf{D}_{\mathbb{R}_+ \rightarrow S}$ denote the space of functions $f: \mathbb{R}_+ \rightarrow S$ that are right continuous with left limits, and endow this space with the Skorokhod J_1 -topology.

The superscript -1 is only used to denote the inverse of a function.

For a measure ξ on \mathbb{R}_+^2 and a ξ -integrable function $f: \mathbb{R}_+^2 \rightarrow \mathbb{R}$, define $\langle f, \xi \rangle := \int_{\mathbb{R}_+^2} f d\xi$. If $\xi = (\xi_1, \dots, \xi_I)$ is a vector of such measures, $\langle f, \xi \rangle := (\langle f, \xi_1 \rangle, \dots, \langle f, \xi_I \rangle)$. Let \mathbf{M} be the space of finite nonnegative Borel measures on \mathbb{R}_+^2 endowed with the weak topology: $\xi^k \xrightarrow{w} \xi$ in \mathbf{M} as $k \rightarrow \infty$ if and only if $\langle f, \xi^k \rangle \rightarrow \langle f, \xi \rangle$ for all continuous bounded function $f: \mathbb{R}_+^2 \rightarrow \mathbb{R}$. Weak convergence of elements of \mathbf{M} is equivalent to convergence in the Prokhorov metric: for $\xi, \varphi \in \mathbf{M}$, define

$$d(\xi, \varphi) := \inf \left\{ \varepsilon: \xi(B) \leq \varphi(B^\varepsilon) + \varepsilon \text{ and } \varphi(B) \leq \xi(B^\varepsilon) + \varepsilon \text{ for all nonempty closed } B \subseteq \mathbb{R}_+^2 \right\},$$

where $B^\varepsilon = \{x \in \mathbb{R}_+^2: \inf_{y \in B} \|x - y\| < \varepsilon\}$.

For $\xi, \varphi \in \mathbf{M}^I$, define

$$d_I(\xi, \varphi) := \max_{1 \leq i \leq I} d(\xi_i, \varphi_i).$$

Equipped with the metric $d_I(\cdot, \cdot)$, the space \mathbf{M}^I is separable and complete.

2. Stochastic model. This section contains a detailed description of the model under consideration. In particular, it specifies the structure of the network, the policy it operates under, and the stochastic dynamical assumptions. Also, a stochastic process is introduced that keeps track of the state of the network; see the state descriptor paragraph.

Network structure. Consider a network that consists of a finite number of links labeled by $j = 1, \dots, J$. Traffic offered to the network is represented by elastic flows coming from a finite number of classes labeled by $i = 1, \dots, I$. All class i flows are transferred through a certain subset of links, we call it *route i* . Transfer of a flow starts immediately upon its arrival and is continuous with all links on the route of the flow being traversed simultaneously. Let A be the $J \times I$ incidence matrix, where $A_{ji} = 1$ if route i contains link j and $A_{ji} = 0$ otherwise.

Suppose that at a particular time t the population of the network is $z \in \mathbb{Z}_+^I$, where z_i stands for the number of flows on route i . All flows on route i are transferred at the same rate $\lambda_i(z)$ that is at most $m_i \in (0, \infty)$. If $z_i = 0$, put $\lambda_i(z) := 0$. We refer to $\Lambda_i(z) := \lambda_i(z)z_i$ as the *bandwidth allocated to route i* . The sum of the bandwidths allocated to the routes that contain link j is the *bandwidth allocated through link j* and is at most $C_j \in (0, \infty)$. We call C_j the *capacity of link j* . Hence, the vectors $\lambda(z) = (\lambda_1(z), \dots, \lambda_I(z))$ and $\Lambda(z) = (\Lambda_1(z), \dots, \Lambda_I(z))$ must satisfy

$$A(\lambda(z) * z) = \Lambda(z) \leq C, \quad \lambda(z) \leq m,$$

where $C = (C_1, \dots, C_J)$ and $m = (m_1, \dots, m_I)$ are the vectors of link capacities and rate constraints.

Bandwidth-sharing policy. At each point in time, the link capacities should be distributed among the routes in such a way that the network utility is maximized. Namely, to each flow on route i we assign a utility $\mathcal{U}_i(\cdot)$ that is a function of the rate allocated to that flow. Assume that the functions $\mathcal{U}_i(\cdot)$ are strictly increasing and concave in \mathbb{R}_+ , and twice differentiable in $(0, \infty)$ with $\lim_{x \downarrow 0} \mathcal{U}'_i(x) = \infty$. We also allow $\lim_{x \downarrow 0} \mathcal{U}_i(x) = -\infty$ as, for example, in the case of a logarithmic function. Then, for $z \in \mathbb{R}_+^I$, the vector $\lambda(z)$ of rates is the unique optimal solution to

$$\text{maximize } \sum_{i=1}^I z_i \mathcal{U}_i(\lambda_i) \quad \text{subject to } A(\lambda * z) \leq C, \quad \lambda \leq m, \quad (1)$$

where, by convention, $0 \times (-\infty) := 0$. Although the population vector has integer-valued coordinates, we assume that $\lambda(z)$ and $\Lambda(z) := \lambda(z) * z$ are defined via (1) in the entire orthant \mathbb{R}_+^I to accommodate fluid analogues of the population process later.

The utility maximization procedure (1) implies that $\lambda_i(z) = \Lambda_i(z) = 0$ if $z_i = 0$. The assumption $\lim_{x \downarrow 0} \mathcal{U}'_i(x) = \infty$ guarantees nonidling, that is $\lambda_i(z), \Lambda_i(z) > 0$ if $z_i > 0$. Reed and Zwart [25] proved that the functions $\lambda(\cdot)$

and $\Lambda(\cdot)$ are differentiable in any direction and, in particular, locally Lipschitz continuous in the interior of \mathbb{R}_+^I . We also show continuity of $\Lambda(\cdot)$ on the boundary of \mathbb{R}_+^I (see the appendix).

LEMMA 1. *The bandwidth allocation function $\Lambda(\cdot)$ is continuous in \mathbb{R}_+^I .*

Stochastic assumptions. All stochastic primitives introduced in this paragraph are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator \mathbb{E} .

Suppose at time zero there is an a.s. finite number of flows in the network, we call them *initial flows*. A random vector $Z^0 \in \mathbb{R}_+^I$ represents the initial population, and Z_i^0 is the number of initial flows on route i . New flows arrive to the network according to a stochastic process $E(\cdot) = (E_1(\cdot), \dots, E_I(\cdot))$ with sample paths in the Skorokhod space $\mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbb{R}_+^I}$. The coordinates of the arrival process are independent counting processes. Recall that a *counting process* is a nondecreasing nonnegative integer-valued process starting from zero. For $t \geq 0$, $E_i(t)$ represents the number of flows that have arrived to route i during the time interval $(0, t]$. The k th such arrival occurs at time $U_{ik} = \inf\{t \geq 0: E_i(t) \geq k\}$, it is called *flow k on route i* , $k \in \mathbb{N}$. Simultaneous arrivals are allowed.

Flows abandon the network because of transfer completions or because they run out of patience, depending on what happens earlier for each particular flow. Flow sizes and patience times are drawn from sequences $\{(B_{il}^0, D_{il}^0)\}_{l \in \mathbb{N}}$, $\{(B_{ik}, D_{ik})\}_{k \in \mathbb{N}}$, $i = 1, \dots, I$, of $(0, \infty)^2$ -valued r.v.'s. For $l = 1, \dots, Z_i^0$, B_{il}^0 and D_{il}^0 represent the residual size and residual patience time at time zero of initial flow l on route i . For $k \in \mathbb{N}$, B_{ik} and D_{ik} represent the initial size and initial patience time of flow k on route i , where “initial” means as upon arrival at time U_{ik} . Let (B_{ik}, D_{ik}) , $k \in \mathbb{N}$, be i.i.d. copies of a r.v. (B_i, D_i) with distribution law θ_i ; and let the mean values $\mathbb{E}B_i =: 1/\mu_i$ and $\mathbb{E}D_i = 1/\nu_i$ be finite. Assume that the sequences $\{(B_{ik}, D_{ik})\}_{k \in \mathbb{N}}$ are independent and do not depend on the arrival process $E(\cdot)$. For the moment, we do not make any specific assumptions about the sequences $\{(B_{il}^0, D_{il}^0)\}_{l \in \mathbb{N}}$.

State descriptor. We denote the *population process* by $Z(\cdot) = (Z_1(\cdot), \dots, Z_I(\cdot))$, where $Z_i(t)$ is the number of flows on route i at time t . As can be seen from what follows, $Z(\cdot)$ is a random element of the Skorokhod space $\mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbb{R}_+^I}$.

For $i = 1, \dots, I$, introduce operators $S_i: \mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbb{R}_+^I} \rightarrow \mathbf{C}_{\mathbb{R}_+^2 \rightarrow \mathbb{R}_+}$ defined by

$$S_i(z, s, t) := \int_s^t \lambda_i(z(u)) du.$$

For $t \geq s \geq 0$, $S_i(Z, s, t)$ is the *cumulative bandwidth allocated per flow on route i during time interval $[s, t]$* . The *residual size and residual lead time at time t* of initial flow $l = 1, \dots, Z_i^0$ on route i are given by

$$B_{il}^0(t) := (B_{il}^0 - S_i(Z, 0, t))^+ \quad \text{and} \quad D_{il}^0(t) := (D_{il}^0 - t)^+,$$

and those of flow $k = 1, \dots, E_i(t)$ on route i by

$$B_{ik}(t) := (B_{ik} - S_i(Z, U_{ik}, t))^+ \quad \text{and} \quad D_{ik}(t) := (D_{ik} - (t - U_{ik}))^+.$$

The state of the network at any time t is defined by the residual sizes and residual patience times of the flows present in the network. With each flow, we associate a dot in \mathbb{R}_+^2 , whose coordinates are the residual size and residual patience time of the flow (see Figure 1). As a flow is getting transferred, the corresponding dot moves toward the axis: to the left at the transfer rate (which is $\lambda_i(Z(t))$ for a flow on route i) and downward at the constant rate of 1. As a dot hits the vertical axis, the corresponding flow leaves because of completion of its transfer. As a dot hits the horizontal axis, the corresponding flow leaves because of impatience. We combine these moving dots into the stochastic process $\mathcal{Z}(\cdot) \in \mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbf{M}^I}$ with

$$\mathcal{Z}_i(t) := \sum_{l=1}^{Z_i^0} \delta_{(B_{il}^0(t), D_{il}^0(t))}^+ + \sum_{k=1}^{E_i(t)} \delta_{(B_{ik}(t), D_{ik}(t))}^+, \tag{2}$$

where, for $x \in \mathbb{R}_+^2$, $\delta_x^+ \in \mathbf{M}$ is the Dirac measure at x if $x_1 \wedge x_2 > 0$ and zero measure otherwise (i.e., assigns a zero mass to any Borel subset of \mathbb{R}_+^2). That is, $\mathcal{Z}_i(t)$ is a counting measure on \mathbb{R}_+^2 that puts a unit mass to each of the dots representing class i flows except those on the axes. The process $\mathcal{Z}(\cdot)$ given by (2) is called the *state descriptor*. Note that the total mass of the state descriptor coincides with the network population, $\langle 1, \mathcal{Z}(\cdot) \rangle = Z(\cdot)$.

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

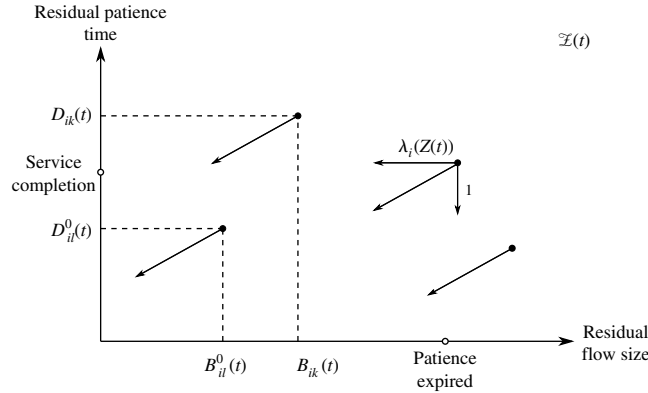


FIGURE 1. The i th coordinate $\mathcal{Z}_i(\cdot)$ of the state descriptor puts a unit mass to the dots representing class i flows except those on the axes.

When proving the results of the paper, we decompose the state descriptors into two parts keeping track of initial and newly arriving flows, respectively. That is,

$$\mathcal{Z}(\cdot) = \mathcal{Z}^{\text{init}}(\cdot) + \mathcal{Z}^{\text{new}}(\cdot),$$

where

$$\mathcal{Z}_i^{\text{init}}(t) := \sum_{l=1}^{Z_i(0)} \delta_{(B_{il}^0(t), D_{il}^0(t))}^+ \quad \text{and} \quad \mathcal{Z}_i^{\text{new}}(t) := \sum_{k=1}^{E_i(t)} \delta_{(B_{ik}(t), D_{ik}(t))}^+.$$

We also define the corresponding total mass processes

$$Z^{\text{init}}(\cdot) := \langle 1, \mathcal{Z}^{\text{init}}(\cdot) \rangle \quad \text{and} \quad Z^{\text{new}}(\cdot) := \langle 1, \mathcal{Z}^{\text{new}}(\cdot) \rangle.$$

3. Fluid model. In this section we define and investigate a fluid model that is a deterministic analogue of the stochastic model described in the previous section. Later on the fluid model will be shown to arise as the limit of the stochastic model under a proper scaling. This convergence implies, in particular, existence of the fluid model.

To define the fluid model we need data $(\eta, \theta, \zeta^0) \in (0, \infty)^I \times \mathbf{M}^I \times \mathbf{M}^I$. The coordinates of η play the role of arrival rates. As in the previous section, θ_i is the joint distribution of the generic size B_i and patience time D_i of a newly arrived flow on route i with finite expectations $\mathbb{E}B_i = 1/\mu_i$ and $\mathbb{E}D_i = 1/\nu_i$. We also introduce the constants

$$\rho_i := \eta_i/\mu_i, \quad \sigma_i := \eta_i/\nu_i,$$

and the vectors $\rho, \sigma \in (0, \infty)^I$,

$$\rho := (\rho_1, \dots, \rho_I), \quad \sigma := (\sigma_1, \dots, \sigma_I).$$

Finally, the measure-valued vector ζ^0 characterizes the initial state of the network. Put $z^0 := \langle 1, \zeta^0 \rangle$ and, for all i , take a r.v. (B_i^0, D_i^0) that is degenerate at $(0, 0)$ if $z_i^0 = 0$ and has distribution ζ_i^0/z_i^0 otherwise. Then z^0 represents the initial population, and (B_i^0, D_i^0) the generic size and patience time of an initial flow on route i . We only consider initial conditions ζ^0 such that the (marginal) distributions of B_i^0 and D_i^0 have no atoms. This restriction is necessary because we require the fluid model to be continuous; see Definition 1.

Denote by \mathcal{C} the collection of corner sets,

$$\mathcal{C} := \{[x, \infty) \times [y, \infty) : (x, y) \in \mathbb{R}_+^2\}.$$

DEFINITION 1. A pair $(\zeta, z) \in \mathbf{C}_{\mathbb{R}_+ \rightarrow \mathbf{M}^I} \times \mathbf{C}_{\mathbb{R}_+ \rightarrow \mathbb{R}_+^I}$ is called a *fluid model solution (FMS)* for the data (η, θ, ζ^0) if $z(\cdot) = \langle 1, \zeta(\cdot) \rangle$ and, for all $i, t \geq 0$ and $A \in \mathcal{C}$,

$$\begin{aligned} \zeta_i(t)(A) &= z_i^0 \mathbb{P}\{(B_i^0, D_i^0) \in A + (S_i(z, 0, t), t)\} \\ &\quad + \eta_i \int_0^t \mathbb{P}\{(B_i, D_i) \in A + (S_i(z, s, t), t - s)\} ds. \end{aligned} \tag{3}$$

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53 . For personal use only, all rights reserved.

In particular, for all i and $t \geq 0$,

$$z_i(t) = \zeta_i(t)(\mathbb{R}_+^2) = z_i^0 \mathbb{P}\{B_i^0 \geq S_i(z, 0, t), D_i^0 \geq t\} + \eta_i \int_0^t \mathbb{P}\{B_i \geq S_i(z, s, t), D_i \geq t - s\} ds. \quad (4)$$

The function $\zeta(\cdot)$ is called a *measure-valued fluid model solution (MVFMS)* and the function $z(\cdot)$ a *numeric fluid model solution (NFMS)*.

Equations (3) and (4) have appealing physical interpretations. For example, (4) simply means that a flow is still in the network at time t if its size and patience exceed, respectively, the amount of service it has received and the time that has passed since its arrival up to time t .

REMARK 1. By Dynkin’s π - λ theorem (see Gromoll et al. [15, §2.3]), FMSs satisfy (3) with any Borel set $A \subseteq \mathbb{R}_+^2$.

REMARK 2. FMSs are invariant with respect to time shifts in the sense that, if $(\zeta, z)(\cdot)$ is an FMS, then, for any $\delta > 0$, $(\zeta^\delta, z^\delta)(\cdot) := (\zeta, z)(\cdot + \delta)$ is an FMS for the data $(\eta, \theta, \zeta(\delta))$. That is, for all i , $t \geq \delta$ and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i(t)(A) = \zeta_i(\delta)(A + (S_i(z, \delta, t), t - \delta)) + \eta_i \int_\delta^t \mathbb{P}\{(B_i, D_i) \in A + (S_i(z, s, t), t - s)\} ds, \quad (5a)$$

$$z_i(t) = \zeta_i(\delta)([S_i(z, \delta, t), \infty) \times [t - \delta, \infty)) + \eta_i \int_\delta^t \mathbb{P}\{B_i \geq S_i(z, s, t), D_i \geq t - s\} ds. \quad (5b)$$

REMARK 3. The measure-valued and numeric components of an FMS uniquely define each other. In particular, uniqueness of an NFMS implies uniqueness of an MVFMS, and the other way around.

As was mentioned earlier, the existence of FMSs is guaranteed by Theorem 5 that follows in the next section. In the rest of this section, we discuss sufficient conditions for an FMS to be unique and for an invariant (i.e., constant) FMS to be unique and asymptotically stable. To prove the stability result, we derive relations for asymptotic bounds for FMSs, which seems to be a novel approach since we have not seen analogous results in the related literature. We also give an example of multiple invariant FMSs.

Uniqueness of an FMS. The proof of the following theorem follows along the lines of the proofs of similar results (Borst et al. [6, Proposition 4.2] and Gromoll et al. [15, Theorem 3.5]); see §6.

THEOREM 1. Suppose that either (i) $z_i^0 = 0$ for all i , or (ii) $z^0 \in (0, \infty)^I$ and the first projection of ζ^0 is Lipschitz continuous, i.e., there exists a constant $L \in (0, \infty)$ such that for all i , $x < x'$ and y ,

$$\zeta_i^0([x, x'] \times [y, \infty)) \leq L(x' - x).$$

Then an FMS for the data (η, θ, ζ^0) is unique.

Uniqueness of an invariant FMS. Let (ζ^*, z^*) be an invariant FMS. By Lemma 3 in §6, all of the coordinates of z^* are positive, and the fluid model Equations (3) and (4) for (ζ^*, z^*) look as follows: for all i , Borel subsets $A \subseteq \mathbb{R}_+^2$ and $t \geq 0$,

$$\zeta_i^*(A) = \zeta_i^*(A + (\lambda_i(z^*)t, t)) + \eta_i \int_0^t \theta_i(A + (\lambda_i(z^*)s, s)) ds, \quad (6)$$

$$z_i^* = \zeta_i^*([\lambda_i(z^*)t, \infty) \times [t, \infty)) + \eta_i \int_0^t \mathbb{P}\{B_i \geq \lambda_i(z^*)s, D_i \geq s\} ds. \quad (7)$$

Letting $t \rightarrow \infty$ in (6) and (7), we obtain the equations

$$\zeta_i^*(A) = \eta_i \int_0^\infty \theta_i(A + (\lambda_i(z^*)s, s)) ds, \quad (8)$$

$$z_i^* = \eta_i \mathbb{E}(B_i / \lambda_i(z^*) \wedge D_i), \quad (9)$$

which are actually equivalent to (6) and (7).

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

Thus, we have the closed-form Equation (9) for the numeric components of invariant FMSs, and the corresponding measure-valued components are defined by (8). In particular, uniqueness of an invariant FMS is equivalent to uniqueness of a solution to (9).

Multiplying the coordinates of (9) by the corresponding rates $\lambda_i(z^*)$, we obtain the equivalent equation

$$\Lambda_i(z^*) = g_i(\lambda_i(z^*)) \quad \text{for all } i, \quad (10)$$

where

$$g_i(x) := \eta_i \mathbb{E}(B_i \wedge x D_i), \quad x \geq 0.$$

We suggest a sufficient condition for uniqueness of a solution to (10) (i.e., of an invariant NFMS) that involves the left-most points of supports of certain distributions.

DEFINITION 2. For an \mathbb{R} -valued r.v. X , denote by $\inf X$ the left-most point of its support. Recall that the *support* of X is the minimal (in the sense of inclusion) closed interval S such that $\mathbb{P}\{X \in S\} = 1$.

As we show later (see Lemmas 6 and 7 in §6), if $m_i \leq 1/\inf(D_i/B_i)$, where $1/0 := \infty$ by definition, the function $g_i(\cdot)$ is continuous and strictly increasing in the interval $[0, m_i]$, implying that its inverse is well defined in $[0, g_i(m_i)]$. Then we can prove the following result.

THEOREM 2. *Let*

$$\inf(D_i/B_i) \leq 1/m_i \quad \text{for all } i. \quad (11)$$

Then there exists a unique invariant FMS (ζ^, z^*) , and the bandwidth allocation vector $\Lambda(z^*)$ is the unique solution to the optimization problem*

$$\text{maximize } \sum_{i=1}^I G_i(\Lambda_i) \quad \text{subject to } A\Lambda \leq C, \quad \Lambda_i \leq g_i(m_i) \quad \text{for all } i, \quad (12)$$

with strictly concave functions $G_i(\cdot)$ such that $G'_i(\cdot) = \mathcal{U}'_i(g_i^{-1}(\cdot))$ in $[0, g_i(m_i)]$.

REMARK 4. Note that it is realistic to assume that flows do not abandon if they are always served at the maximum rate, i.e., that $D_i \geq B_i/m_i$. For such routes, we have $g_i(m_i) = \rho_i$, and the sufficient uniqueness condition (11) reads as $\inf(D_i/B_i) = 1/m_i$.

The complete proof of Theorem 2 is postponed to §6, here we only discuss the main ideas. As we plug the fixed point Equation (10) into the optimization problem (1) for the rate vector $\lambda(z^*)$, the problem (12) follows through two applications of the Karush-Kuhn-Tucker (KKT) conditions—necessary and sufficient conditions for optimality. The problem (12) is strictly concave and does not depend on z^* . Hence, $\Lambda(z^*) =: \Lambda^*$ is the same for all invariant points z^* . This idea of transforming the optimization problem defining the rate vector combined with the fixed point equation into an independent problem for the bandwidth allocation vector we adopted from Borst et al. [6, Lemma 5.2]. Now, since the functions $g_i(\cdot)$ are invertible in the feasible rate intervals $[0, m_i]$, it follows from (10) that the fixed point is unique and given by

$$z_i^* = \Lambda_i^*/g_i^{-1}(\Lambda_i^*). \quad (13)$$

Note that this method not only proves uniqueness of an invariant FMS, but also suggests a two-step algorithm to compute it: first we need to solve the strictly concave optimization problem (12) for Λ^* , which can be done with any desired accuracy in a polynomial time, and then we can compute the fixed point z^* itself by formula (13).

Asymptotic bounds for FMSs. Here we derive asymptotic bounds for NFMSs that, for a wide class of bandwidth-sharing networks, imply convergence to the invariant NFMS provided it is unique.

THEOREM 3. *There exist constants $l, u \in (0, \infty)^I$ such that, for any NFMS $z(\cdot)$,*

$$0 < l_i \leq \underline{\lim}_{t \rightarrow \infty} z_i(t) \leq \overline{\lim}_{t \rightarrow \infty} z_i(t) \leq u_i \quad \text{for all } i.$$

These constants satisfy the relations

$$l_i = \eta_i \mathbb{E}(B_i/R_i(l, u) \wedge D_i), \quad u_i = \eta_i \mathbb{E}(B_i/r_i(l, u) \wedge D_i) \quad \text{for all } i, \quad (14)$$

where the functions $r(\cdot, \cdot)$ and $R(\cdot, \cdot)$ are defined by

$$r_i(x, x') := \inf_{x \leq z \leq x'} \lambda_i(z), \quad R_i(x, x') := \sup_{x \leq z \leq x'} \lambda_i(z) \quad \text{for all } i \text{ and } x \leq x'.$$

REMARK 5. There could be more than one pair (l, u) solving (14). The asymptotic bounds l and u for NFMSs given by Theorem 3 form one of such pairs.

PROOF OF THEOREM 3. Note that if

$$0 < \tilde{l}_i \leq \underline{\lim}_{t \rightarrow \infty} z_i(t) \leq \overline{\lim}_{t \rightarrow \infty} z_i(t) \leq \tilde{u}_i \quad \text{for all } i, \tag{15}$$

then also

$$\eta_i \mathbb{E}(B_i/R_i(\tilde{l}, \tilde{u}) \wedge D_i) \leq \underline{\lim}_{t \rightarrow \infty} z_i(t) \leq \overline{\lim}_{t \rightarrow \infty} z_i(t) \leq \eta_i \mathbb{E}(B_i/r_i(\tilde{l}, \tilde{u}) \wedge D_i) \quad \text{for all } i. \tag{16}$$

Indeed, by (15), for any $\varepsilon \in (0, \min_{1 \leq i \leq I} \tilde{l}_i)$, there exists a t_ε such that

$$\tilde{l}_i - \varepsilon \leq z_i(t) \leq \tilde{u}_i + \varepsilon \quad \text{for all } i \text{ and } t \geq t_\varepsilon.$$

Introduce the vectors

$$\tilde{l} - \varepsilon := (\tilde{l}_1 - \varepsilon, \dots, \tilde{l}_I - \varepsilon), \quad \tilde{u} + \varepsilon := (\tilde{u}_1 + \varepsilon, \dots, \tilde{u}_I + \varepsilon).$$

Then

$$r_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t - s) \leq S_i(z, s, t) \leq R_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t - s) \quad \text{for } t \geq s \geq t_\varepsilon,$$

which, when plugged into the shifted fluid model Equation (5b), implies that

$$\begin{aligned} z_i(t) &\geq \eta_i \int_{t_\varepsilon}^t \mathbb{P}\{B_i \geq R_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t - s), D_i \geq (t - s)\} ds, \\ z_i(t) &\leq \zeta_i(t_\varepsilon)([S_i(z, t_\varepsilon, t), \infty) \times [t - t_\varepsilon, \infty)) \\ &\quad + \eta_i \int_{t_\varepsilon}^t \mathbb{P}\{B_i \geq r_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t - s), D_i \geq (t - s)\} ds \quad \text{for } t \geq t_\varepsilon, \end{aligned}$$

where $\zeta(\cdot)$ is the corresponding MVFMS. Taking $t \rightarrow \infty$ in the last two inequalities, we obtain

$$\eta_i \mathbb{E}(B_i/R_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon) \wedge D_i) \leq \underline{\lim}_{t \rightarrow \infty} z_i(t) \leq \overline{\lim}_{t \rightarrow \infty} z_i(t) \leq \eta_i \mathbb{E}(B_i/r_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon) \wedge D_i),$$

and then (16) follows as $\varepsilon \rightarrow 0$.

Now we will iterate (15)–(16). The rate constraints plugged into (4) imply the initial bounds

$$0 < l_i^0 := \eta_i \mathbb{E}(B_i/m_i \wedge D_i) \leq \underline{\lim}_{t \rightarrow \infty} z_i(t) \leq \overline{\lim}_{t \rightarrow \infty} z_i(t) \leq \eta_i \mathbb{E}D_i =: u_i^0 \quad \text{for all } i,$$

and then (15)–(16) yield the recursive bounds

$$\begin{aligned} l_i^k &:= \eta_i \mathbb{E}(B_i/R_i(l^{k-1}, u^{k-1}) \wedge D_i) \leq \underline{\lim}_{t \rightarrow \infty} z_i(t), \\ u_i^k &:= \eta_i \mathbb{E}(B_i/r_i(l^{k-1}, u^{k-1}) \wedge D_i) \geq \overline{\lim}_{t \rightarrow \infty} z_i(t) \quad \text{for all } k \in \mathbb{N} \text{ and } i. \end{aligned} \tag{17}$$

The sequence $\{l^k\}_{k \in \mathbb{N}}$ is nondecreasing and bounded from above by u^0 . The sequence $\{u^k\}_{k \in \mathbb{N}}$ is nonincreasing and bounded from below by l^0 . Hence, there exist the limits $\lim l^k =: l$ and $\lim u^k =: u$. In (17), let $k \rightarrow \infty$, then (14) follows.

Note finally that the recursive bounds $\{l^k\}_{k \in \mathbb{N}}$ and $\{u^k\}_{k \in \mathbb{N}}$ as well as their limits l and u do not depend on a particular NFMS. \square

Asymptotic stability of an invariant fluid model solution. It is tractable to assume that transfer rates in a bandwidth-sharing network decrease as its population grows. In particular, tree networks satisfy this property; see Borst et al. [6].

DEFINITION 3. If $z' \geq z \in (0, \infty)^I$ implies $\lambda(z') \leq \lambda(z)$, the network is called *monotone*.

For monotone networks, the system of Equation (14) decomposes into two independent systems of equations for the lower bound l and for the upper bound u :

$$l_i = \eta_i \mathbb{E}(B_i / \lambda_i(l) \wedge D_i) \quad \text{for all } i, \quad (18a)$$

$$u_i = \eta_i \mathbb{E}(B_i / \lambda_i(u) \wedge D_i) \quad \text{for all } i, \quad (18b)$$

which implies the following result.

COROLLARY 1. *Suppose that the network is monotone and has a unique invariant FMS (ζ^*, z^*) . Then any FMS $(\zeta, z)(t) \rightarrow (\zeta^*, z^*)$ as $t \rightarrow \infty$.*

Indeed, both (18a) and (18b) coincide with the fixed point Equation (9), and since Corollary 1 assumes that the latter equation has a unique solution z^* , it immediately follows by Theorem 3 that, for any NFMS $z(\cdot)$,

$$l_i = \liminf_{t \rightarrow \infty} z_i(t) = \overline{\lim}_{t \rightarrow \infty} z_i(t) = u_i = z_i^* \quad \text{for all } i.$$

In the appendix we also show that $z(t) \rightarrow z^*$ implies $\zeta(t) \rightarrow \zeta^*$.

Example: Single link. The sufficient condition for uniqueness of an invariant FMS given by Theorem 2 is sometimes also necessary. Consider, for example, processor sharing in critical load, that is $J = I = 1$ and (omitting the link and class indices) $\rho = C$. In this case, the fixed point Equation (9) looks like

$$z^* = \eta \mathbb{E} \left(\frac{B}{C/z^* \wedge m} \wedge D \right),$$

which, for z^* such that $C/z^* \leq m$ and $Bz^*/C \leq D$ a.s., reduces to

$$z^* = \eta \mathbb{E}(Bz^*/C \wedge D) = \eta \mathbb{E}Bz^*/C = \rho z^*/C = z^*.$$

I.e., any $z^* \in [C/m, C \inf D/B]$ is an invariant NFMS. In particular, if $\inf D/B > 1/m$, which violates the assumption of Theorem 2, then there is a continuum of invariant FMSs.

For a single link critically loaded by multiple classes of flows, we have an analogous result, which is more complicated to derive and therefore the proof is postponed to §6.

THEOREM 4. *Assume that $J = 1$ (in what follows we omit the link index), and that the utility functions are $\mathcal{U}_i(x) = \kappa_i \log x$. If $\sum_{i=1}^I \eta_i \mathbb{E}(B_i/m_i \wedge D_i) \neq C$, then there is a unique invariant FMS. Otherwise there might be a continuum of invariant FMSs.*

4. Sequence of stochastic models and fluid limit theorem. In this section we study the asymptotic behavior of the stochastic network described in §2 as its global parameters—capacities and arrival rates—grow large, while the characteristics of an individual flow remain of a fixed order. We refer to this scaling as the large capacity regime.

Large capacity scaling. To a sequence \mathcal{R} of positive numbers increasing to ∞ , we associate a sequence of stochastic models as defined in §2. We mark all parameters associated with the r th model with a superscript r and assume the following:

(A.1) Network structure, rate constraints, and utility function are the same in all models: $A^r = A$, $m^r = m$ and $\mathcal{U}_i^r(\cdot) = \mathcal{U}_i(\cdot)$ for all i .

(A.2) Link capacities grow linearly in r : $C^r = rC$.

(A.3) Arrival rates grow linearly in r : $\bar{E}^r(\cdot) := E^r(\cdot)/r \Rightarrow \eta(\cdot)$ as $r \rightarrow \infty$, where $\eta(t) := t\eta$ and $\eta \in (0, \infty)^I$.

(A.4) Flow sizes and patience times remain of a fixed order: for all i , $(B_i^r, D_i^r) \Rightarrow (B_i, D_i)$ as $r \rightarrow \infty$, where (B_i, D_i) are $(0, \infty)^I$ -valued r.v.'s with distributions θ_i and finite mean values $(1/\mu_i, 1/\nu_i)$, and also $(1/\mu_i^r, 1/\nu_i^r) \rightarrow (1/\mu_i, 1/\nu_i)$.

(A.5) The scaled initial configuration converges in distribution to a random vector of finite measures: $\bar{\mathcal{X}}^r(0) := \mathcal{X}^r(0)/r \Rightarrow \zeta^0$.

(A.6) The projections $\zeta_i^0(\cdot \times \mathbb{R}_+)$ and $\zeta_i^0(\mathbb{R}_+ \times \cdot)$ are a.s. free of atoms for all i .

Fluid limit theorem. In the large capacity regime, the stochastic model defined in §2 converges to the fluid model defined in §3. More precisely, introduce the scaled state descriptors and population processes

$$\tilde{\mathcal{Z}}^r(\cdot) := \mathcal{Z}^r(\cdot)/r, \quad \bar{Z}^r(\cdot) := \langle 1, \tilde{\mathcal{Z}}^r(\cdot) \rangle = Z^r(\cdot)/r.$$

Also introduce the scaled versions of the two components of the state descriptor:

$$\begin{aligned} \tilde{\mathcal{Z}}^{r, \text{init}}(\cdot) &:= \mathcal{Z}^{r, \text{init}}(\cdot)/r, & \bar{Z}^{r, \text{init}}(\cdot) &:= \langle 1, \tilde{\mathcal{Z}}^{r, \text{init}}(\cdot) \rangle = Z^{r, \text{init}}(\cdot)/r, \\ \tilde{\mathcal{Z}}^{r, \text{new}}(\cdot) &:= \mathcal{Z}^{r, \text{new}}(\cdot)/r, & \bar{Z}^{r, \text{new}}(\cdot) &:= \langle 1, \tilde{\mathcal{Z}}^{r, \text{new}}(\cdot) \rangle = Z^{r, \text{new}}(\cdot)/r. \end{aligned}$$

REMARK 6. Let $\lambda(\cdot)$ be the rate allocation function in the unscaled network, then

$$\begin{aligned} \lambda^r(z) &:= \arg \max_{\substack{A(\lambda * z) \leq rC \\ \lambda \leq m}} \sum_{i=1}^I z_i \mathcal{U}_i(\lambda_i) \\ &= \arg \max_{\substack{A(\lambda * z/r) \leq C \\ \lambda \leq m}} \sum_{i=1}^I (z_i/r) \mathcal{U}_i(\lambda_i) =: \lambda(z/r), \end{aligned}$$

and

$$S_i^r(Z^r, s, t) := \int_s^t \lambda_i^r(Z^r(u)) du = \int_s^t \lambda_i(\bar{Z}^r(u)) du =: S_i(\bar{Z}^r, s, t).$$

We now provide the definition of a fluid limit followed by the main result of this section.

DEFINITION 4. We refer to weak limits along convergent subsequences $\{(\tilde{\mathcal{Z}}^q, \bar{Z}^q)(\cdot)\}_{q \in \mathbb{Q}}$, $\mathbb{Q} \subseteq \mathbb{R}$, as *fluid limits*.

THEOREM 5. Under the assumptions (A.1)–(A.6), the sequence $\{(\tilde{\mathcal{Z}}^r, \bar{Z}^r)(\cdot)\}_{r \in \mathbb{R}}$ is tight in $\mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbf{M}^I} \times \mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbb{R}_+^I}$, and all fluid limits are a.s. FMSs for the data (η, θ, ζ^0) . In particular, if there is a unique FMS $(\mathcal{Z}, Z)(\cdot)$ for the data (η, θ, ζ^0) , then $(\tilde{\mathcal{Z}}^r, \bar{Z}^r)(\cdot) \Rightarrow (\mathcal{Z}, Z)(\cdot)$ as $r \rightarrow \infty$.

The proof follows in §7. To show tightness we adjust the techniques of Gromoll and Williams [14] to the two-dimensional case, since in Gromoll and Williams [14] flows are patient and state descriptors are vectors of measures on \mathbb{R}_+ . The proof of convergence to FMSs follows the lines of that in Gromoll et al. [15]. It uses the boundedness of fluid limits away from zero, and the key difference is that in Gromoll et al. [15] this property is guaranteed by the overload regime, whereas in our model it holds in any load regime because of the rate constraints.

5. Convergence of stationary distributions. Assume that, in the stochastic model defined in §2, the arrival processes are Poisson of rates η_1, \dots, η_I . Then there exists a unique stationary (and also limiting as $t \rightarrow \infty$) distribution of the state descriptor $\mathcal{Z}(\cdot)$. Indeed, without loss of generality, there are i.i.d. r.v.'s $\{\tilde{D}_{ik}\}_{k \in \mathbb{N}, 1 \leq i \leq I}$ distributed as $\max_{1 \leq i \leq I} D_i$ and such that a.s. $D_{ik} \leq \tilde{D}_{ik}$ for all k and i . Then the total population $\sum_{i=1}^I Z_i(\cdot)$ of the network is a.s. and within the whole time horizon \mathbb{R}_+ bounded from above by the length of the $M/G/\infty$ queue with the following parameters. At time $t = 0$, there are $\sum_{i=1}^I Z_i(0)$ customers in the queue whose service times are patience times of initial flows in the network. The input process for the queue is the composition of those for the network, and hence is Poisson of rate $\sum_{i=1}^I \eta_i$. Service times of new customers in the queue are drawn from the sequence $\{\tilde{D}_{ik}\}_{k \in \mathbb{N}, 1 \leq i \leq I}$ of upper bounds for patience times of new flows in the network. As any other $M/G/\infty$ queue, the defined queue is regenerative. The instants when a customer enters the empty queue form an embedded renewal process whose cycle length is non-lattice and has a finite mean value $\exp(\sum_{i=1}^I \eta_i \mathbb{E} \tilde{D}_{11}) / \sum_{i=1}^I \eta_i$. With respect to this renewal process, the state descriptor $\mathcal{Z}(\cdot)$ is also regenerative. Then, by Asmussen [1, Chap. VI, Theorem 1.2], there exists a limiting distribution for $\mathcal{Z}(\cdot)$.

Now consider a sequence of stochastic models as defined in §2 that satisfies the assumptions (A.1), (A.2), (A.4) (see §4) and

(A'.3) the input processes $E_1^r(\cdot), \dots, E_I^r(\cdot)$ are independent Poisson processes of rates $\eta_1^r, \dots, \eta_I^r$, and $\eta^r/r \rightarrow \eta \in (0, \infty)^I$ as $r \rightarrow \infty$,

(A'.4) on all routes i , the size B_i^r of a flow and its patience time D_i^r are independent.

Let \mathcal{Y}^r have the stationary distribution of $\mathcal{L}^r(\cdot)$ and put $Y^r := \langle 1, \mathcal{Y}^r \rangle$. Introduce also the scaled versions

$$\bar{\mathcal{Y}}^r := \mathcal{Y}^r / r, \quad \bar{Y}^r := \langle 1, \bar{\mathcal{Y}}^r \rangle = Y^r / r.$$

We now have the following result.

THEOREM 6. *Under the assumptions (A.1), (A.2), (A'.3), (A.4), and (A'.4), the sequence of the fluid scaled stationary distributions $\{(\bar{\mathcal{Y}}^r, \bar{Y}^r)\}_{r \in \mathcal{R}}$ is tight, and any weak limit point (\mathcal{Y}, Y) is a weak invariant FMS, i.e., there exists a stationary FMS $(\mathcal{Z}, Z)(t) \stackrel{d}{=} (\mathcal{Y}, Y), t \geq 0$. In particular, by Corollary 1, if the network is monotone and has a unique invariant FMS (ζ^*, z^*) , then $(\bar{\mathcal{Y}}^r, \bar{Y}^r) \Rightarrow (\zeta^*, z^*)$ as $r \rightarrow \infty$.*

The general strategy of the proof is adopted from [18, Theorem 3.3]: we check that any convergent subsequence of initial conditions $\{\tilde{\mathcal{Z}}^q(0) \stackrel{d}{=} \bar{\mathcal{Y}}^q\}_{q \in \mathcal{Q}}, \mathcal{Q} \subseteq \mathcal{R}, \bar{\mathcal{Y}}^q \Rightarrow \mathcal{Y}$, satisfies the assumptions of the fluid limit theorem (we only need to check (A.6)). Then the corresponding subsequence $\{\tilde{\mathcal{Z}}^q(\cdot)\}_{q \in \mathcal{Q}}$ of the scaled state descriptors converges to an MVFMS that is stationary (i.e., \mathcal{Y} is a weak invariant MVFMS) since all $\tilde{\mathcal{Z}}^q(\cdot)$ are stationary.

The techniques we use to implement this strategy are different from the techniques of [18], though. Our key instruments for establishing tightness are $M/G/\infty$ bounds; see §8. Below we present an elegant proof of (A.6) for weak limit points of $\{\mathcal{Y}^r\}_{r \in \mathcal{R}}$.

LEMMA 2. *Any weak limit point \mathcal{Y} of $\{\bar{\mathcal{Y}}^r\}_{r \in \mathcal{R}}$ has both projections $\mathcal{Y}(\cdot \times \mathbb{R}_+)$ and $\mathcal{Y}(\mathbb{R}_+ \times \cdot)$ a.s. free of atoms.*

PROOF. The key idea is the following. Consider the network in its stationary regime. Then, on one hand, it always has the same distribution, and on the other hand, all initial flows are gone at some point, and newly arriving flows do not accumulate along horizontal and vertical lines.

Let \mathcal{Y} be the weak limit along a subsequence $\{\bar{\mathcal{Y}}^q\}_{q \in \mathcal{Q}}$, and run the q th network starting from $\tilde{\mathcal{Z}}^q(0) \stackrel{d}{=} \bar{\mathcal{Y}}^q$. By Gromoll et al. [15, Lemma 6.2], it suffices to show that, for any $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that

$$\liminf_{q \rightarrow \infty} \mathbb{P}^q \left\{ \sup_{x \in \mathbb{R}_+} \|\bar{\mathcal{Y}}^q(H_x^{x+a})\| \vee \|\bar{\mathcal{Y}}^q(V_x^{x+a})\| \leq \delta \right\} \geq 1 - \varepsilon, \tag{19}$$

where $H_a^b := \mathbb{R}_+ \times [a, b]$ and $V_a^b := [a, b] \times \mathbb{R}_+$ for all $b \geq a \geq 0$.

First we estimate the time when there is only a few (when scaled) initial flows left. The initial flows whose initial patience times are less than t are already gone at time t . Then Lemma 16 (see §8) implies that (recall that $\Pi(\lambda)$ stands for the Poisson distribution with parameter λ)

$$\begin{aligned} \bar{Z}_i^{q, \text{init}}(t) \leq \tilde{\mathcal{Z}}_i^q(0)(\mathbb{R}_+ \times [t, \infty)) &\stackrel{d}{=} \bar{\mathcal{Y}}_i^q(\mathbb{R}_+ \times [t, \infty)) \\ &\leq \frac{1}{q} \Pi \left(\eta_i^q \int_t^\infty \mathbb{P}^q \{D_i^q > y\} dy \right) \Rightarrow \eta_i \int_t^\infty \mathbb{P} \{D_i > y\} dy. \end{aligned}$$

Take T such that $\max_{1 \leq i \leq I} \eta_i \int_t^\infty \mathbb{P} \{D_i > y\} dy < \delta/2$, then

$$\lim_{q \rightarrow \infty} \mathbb{P}^q \{ \|\bar{Z}^{q, \text{init}}(T)\| \leq \delta/2 \} = 1.$$

Now, in Lemma 11 (see §7), we prove that newly arriving customers do not accumulate in thin horizontal and vertical strips, i.e., there exists an $a > 0$ such that

$$\lim_{q \rightarrow \infty} \mathbb{P}^q \left\{ \sup_{t \in [0, T]} \sup_{x \in \mathbb{R}_+} \{ \|\tilde{\mathcal{Z}}^{q, \text{new}}(t)(H_x^{x+a})\| \vee \|\tilde{\mathcal{Z}}^{q, \text{new}}(t)(V_x^{x+a})\| \} \leq \delta/2 \right\} \geq 1 - \varepsilon.$$

Finally, because of stationarity of \mathcal{Y}^q ,

$$\begin{aligned} &\mathbb{P}^q \left\{ \sup_{x \in \mathbb{R}_+} \{ \|\bar{\mathcal{Y}}^q(H_x^{x+a})\| \vee \|\bar{\mathcal{Y}}^q(V_x^{x+a})\| \} \leq \delta \right\} \\ &= \mathbb{P}^q \left\{ \sup_{x \in \mathbb{R}_+} \{ \|\tilde{\mathcal{Z}}^q(T)(H_x^{x+a})\| \vee \|\tilde{\mathcal{Z}}^q(T)(V_x^{x+a})\| \} \leq \delta \right\} \\ &\geq \mathbb{P}^q \left\{ \|\bar{Z}^{q, \text{init}}(T)\| \leq \delta/2, \sup_{x \in \mathbb{R}_+} \{ \|\tilde{\mathcal{Z}}^{q, \text{new}}(T)(H_x^{x+a})\| \vee \|\tilde{\mathcal{Z}}^{q, \text{new}}(T)(V_x^{x+a})\| \} \leq \delta/2 \right\}, \end{aligned}$$

which implies (19) by the choice of T and a . \square

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

6. Proof of fluid model properties. Here we prove the results of §3.

6.1. Proof of Theorem 1. In the proof of Theorem 1, we exploit boundedness of NFMSs away from zero and in the norm (see Lemma 3), and Lipschitz continuity of MVFMSs in the first coordinate (see Lemma 5). We also use the auxiliary Lemma 4.

Recall the notations $\sigma_i = \eta_i \mathbb{E}D_i$ and $\sigma = (\sigma_1, \dots, \sigma_I)$.

LEMMA 3. *Let $z(\cdot)$ be an NFMS. Then $\sup_{t \geq 0} \|z(t)\| \leq \|z(0)\| + \|\sigma\| < \infty$ and, for any $\delta > 0$, $\inf_{t \geq \delta} \min_{1 \leq i \leq I} z_i(t) > 0$. In particular, if $z_i(0) > 0$, then $\inf_{t \geq 0} z_i(t) > 0$.*

PROOF. By the rate constraints, $S_i(z, s, t) \leq m_i(t - s)$ for all $s \leq t$, which, when plugged into the fluid model Equation (4), implies the following lower bound: $z_i(t) \geq \eta_i \int_0^t \mathbb{P}\{B_i/m_i \wedge D_i \geq s\} ds$. Since $f_i(s) := \mathbb{P}\{B_i/m_i \wedge D_i \geq s\} \uparrow \mathbb{P}\{B_i/m_i \wedge D_i > 0\} = 1$ as $s \downarrow 0$, in a small enough interval $[0, \varepsilon]$, $f_i(\cdot) \geq 1/2$. Then, for $t \geq \delta$, $z_i(t) \geq \eta_i \int_0^{\delta \wedge \varepsilon} f_i(s) ds \geq \eta_i(\delta \wedge \varepsilon)/2$. The upper bound follows from (4) directly: $z_i(t) \leq z_i(0) + \eta_i \int_0^t \mathbb{P}\{D_i \geq s\} ds \uparrow z_i(0) + \sigma_i$ as $t \uparrow \infty$. \square

LEMMA 4. *For an \mathbb{R} -valued r.v. ξ and $x \leq x'$, $\int_{\mathbb{R}} \mathbb{P}\{u + x \leq \xi \leq u + x'\} du \leq x' - x$.*

See the appendix for the proof.

LEMMA 5. *Under assumption (ii) of Theorem 1, any MVFMS $\zeta(\cdot)$ at any time $t \geq 0$ has a Lipschitz continuous first projection, i.e., there exists a constant $L(\zeta, t) \in (0, \infty)$ such that for all i , $x < x'$ and y ,*

$$\zeta_i(t)([x, x'] \times [y, \infty)) \leq L(\zeta, t)(x' - x).$$

PROOF. For an FMS $(\zeta, z)(\cdot)$, for all i , $t \geq 0$, $x < x'$ and y ,

$$\zeta_i(t)([x, x'] \times [y, \infty)) \leq f_i(x, x', y) + \eta_i g_i(x, x', y),$$

where

$$\begin{aligned} f_i(x, x', y) &:= \zeta_i^0([x + S_i(z, 0, t), x' + S_i(z, 0, t)] \times [y + t, \infty)), \\ g_i(x, x', y) &:= \int_0^t \mathbb{P}\{x + S_i(z, s, t) \leq B_i \leq x' + S_i(z, s, t)\} ds. \end{aligned} \tag{20}$$

By Lipschitz continuity of the initial condition, $f_i(x, x', y) \leq L(x' - x)$. In (20), change the variable of integration for $v = V(s) := S_i(z, s, t)$. Then

$$g_i(x, x', y) = \int_0^{S_i(z, 0, t)} \mathbb{P}\{x + v \leq B_i \leq x' + v\} / \lambda_i(z(V^{-1}(v))) dv \leq M(\zeta, t)(x' - x),$$

where $M(\zeta, t) := \sup_{s \in [0, t]} \max_{1 \leq i \leq I} 1/\lambda_i(z(s))$. By Lemma 3, the functions $1/\lambda_i(z(\cdot))$ are continuous in $[0, t]$. Hence $M(\zeta, t)$ is finite and the first projection of $\zeta(t)$ is Lipschitz continuous with the constant $L(\zeta, t) := L + \|\eta\|M(\zeta, t)$. \square

Now we are in a position to prove Theorem 1.

PROOF OF THEOREM 1. Let $(\zeta^1, z^1)(\cdot)$ and $(\zeta^2, z^2)(\cdot)$ be two FMSs for the data (η, θ, ζ^0) .

(i) We show that the two FMSs coincide in an interval $[0, \delta]$. We check that $z^1(\delta) = z^2(\delta) \in (0, \infty)^I$ and that the first projection of $\zeta^1(\delta) = \zeta^2(\delta)$ is Lipschitz continuous. Then, by Remark 2 and the second part of the theorem, the two FMSs coincide everywhere.

Note that, for a vector $z \in (0, \infty)^I$ of a small enough norm, $\lambda_i(z) = m_i$ for all i . Lemma 3 and the fluid model Equation (4) imply that $0 < z_i^1(t), z_i^2(t) \leq \eta_i t$ for all i and $t > 0$. Then, for all i and $s, t \in [0, \delta]$, where δ is small enough,

$$S_i(z^1, s, t) = S_i(z^2, s, t) = m_i(t - s). \tag{21}$$

Plugging (21) into (4), we obtain, for $t \in [0, \delta]$ and all i ,

$$z_i^1(t) = z_i^2(t) = \eta_i \int_0^t \mathbb{P}\{B_i/m_i \wedge D_i \geq s\} ds.$$

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

By Remark 3, $\zeta^1(\cdot)$ and $\zeta^2(\cdot)$ coincide in $[0, \delta]$, too. Lipschitz continuity of the first projection of $\zeta^1(\delta) = \zeta^2(\delta)$ follows as we plug (21) into the fluid model Equation (3) (recall that it is valid for all Borel sets): for all i , $x < x'$ and y ,

$$\begin{aligned} \zeta_i^j(\delta)([x, x'] \times [y, \infty)) &= \eta_i \int_0^\delta \mathbb{P}\{x + m_i s \leq B_i \leq x' + m_i s, D_i \geq y + s\} ds \\ &\leq \eta_i \int_0^\delta \mathbb{P}\{x/m_i + s \leq B_i/m_i \leq x'/m_i + s\} ds \\ &\leq \eta_i(x' - x)/m_i, \quad j = 1, 2, \end{aligned}$$

where the last inequality holds by Lemma 4.

(ii) Suppose that the two FMSs are different, that is $t_* := \inf\{t > 0: z^1(t) \neq z^2(t)\} < \infty$.

Without loss of generality we may assume that $t_* = 0$. Indeed, otherwise we can consider the time-shifted FMSs $(\zeta^j, z^j)(t_* + \cdot)$, $j = 1, 2$. By Lemmas 3 and 5, they start from $z^1(t_*) = z^2(t_*) \in (0, \infty)^l$ and $\zeta^1(t_*) = \zeta^2(t_*)$ with a Lipschitz continuous first projection.

By Lemma 3, the two NFMS never leave a compact set $[\delta, \Delta]^l \subset (0, \infty)^l$. Since the rate functions $\lambda_i(z)$ are Lipschitz continuous in such sets, there exists a constant $K \in (0, \infty)$ such that, for all i and $s \leq t$,

$$|S_i(z^1, s, t) - S_i(z^2, s, t)| \leq Kt \sup_{s \in [0, t]} \|z^1(s) - z^2(s)\| =: Kt\varepsilon(t).$$

Then, by Lipschitz continuity of the initial condition, we have, for all i and $t \geq 0$,

$$|z_i^1(t) - z_i^2(t)| \leq LKt\varepsilon(t) + \eta_i \int_0^t \mathbb{P}\{S_i(z^1, s, t) - Kt\varepsilon(t) \leq B_i \leq S_i(z^1, s, t) + Kt\varepsilon(t)\} ds.$$

In the last equation, change the variable of integration for $v = S_i(z^1, s, t)$ (cf. the proof of Lemma 5) and put $M = \sup_{z \in [\delta, \Delta]^l} \max_{1 \leq i \leq l} 1/\lambda_i(z)$. Then

$$|z_i^1(t) - z_i^2(t)| \leq LKt\varepsilon(t) + \eta_i M 2Kt\varepsilon(t) \quad \text{for all } i \text{ and } \varepsilon(t) \leq (L + 2\|\eta\|M)Kt\varepsilon(t),$$

which implies that $\varepsilon(t) = 0$ for small enough t , and we arrive at a contradiction with $t_* = 0$. \square

6.2. Proof of Theorem 2. Before proceeding with the proof of the theorem, we discuss some properties of the functions $g_i(\cdot)$ in the auxiliary Lemmas 6 and 7. Recall that these functions are given by

$$g_i(x) = \eta_i \mathbb{E}(B_i \wedge xD_i), \quad x \geq 0.$$

LEMMA 6. *The function $g_i(\cdot)$ is continuous. Also $g_i(\cdot)$ is strictly increasing in $[0, \alpha_i]$ and constant in $[\alpha_i, \infty)$, where*

$$\alpha_i := \inf\{x: g_i(x) = \rho_i\} > 0,$$

and infimum over the empty set is defined to be ∞ .

PROOF. Continuity of $g_i(\cdot)$ follows by the dominated convergence theorem.

The situation $\alpha_i = 0$ is not possible since in that case $g_i(x) = \rho_i$ for all $x > 0$ by the definition of α_i . But $g_i(\cdot)$ is continuous and $g_i(x) \rightarrow g_i(0) = 0$ as $x \rightarrow 0$.

If $\alpha_i < \infty$, then, again by the definition of α_i and continuity of $g_i(\cdot)$, we have $g_i(x) = \rho_i$ for all $x \geq \alpha_i$ and $g_i(x) < \rho_i = g_i(\alpha_i)$ for all $x < \alpha_i$.

It is left to check that $g_i(\cdot)$ is strictly increasing in $[0, \alpha_i)$. Assume that $0 \leq x < y < \alpha_i$, but $g_i(x) = g_i(y)$. Then

$$\begin{aligned} 0 = g_i(y)/\eta_i - g_i(x)/\eta_i &= \mathbb{E}B_i \mathbb{1}_{\{B_i \leq xD_i\}} + \mathbb{E}B_i \mathbb{1}_{\{xD_i < B_i \leq yD_i\}} + \mathbb{E}yD_i \mathbb{1}_{\{B_i > yD_i\}} \\ &\quad - \mathbb{E}B_i \mathbb{1}_{\{B_i \leq xD_i\}} - \mathbb{E}xD_i \mathbb{1}_{\{xD_i < B_i \leq yD_i\}} - \mathbb{E}xD_i \mathbb{1}_{\{B_i > yD_i\}} \\ &= \mathbb{E} \underbrace{(B_i - xD_i) \mathbb{1}_{\{xD_i < B_i \leq yD_i\}}}_{=: X} + (y - x) \mathbb{E} \underbrace{D_i \mathbb{1}_{\{B_i > yD_i\}}}_{=: Y}, \end{aligned}$$

where the r.v.'s X and Y are nonnegative, so they must a.s. equal zero. In particular, we have $B_i \leq yD_i$ and $g_i(y) = \rho_i$, which contradicts the definition of α_i since $y < \alpha_i$. \square

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

The stabilization points α_i of the functions $g_i(\cdot)$ are related with the r.v.'s (B_i, D_i) in the following way.

LEMMA 7. *If $\alpha_i < \infty$, then $\inf D_i/B_i = 1/\alpha_i$. If $\alpha_i = \infty$, then $\inf D_i/B_i = 0$.*

PROOF. First assume $\alpha_i < \infty$. Rewrite the relation $g_i(x) = \rho_i$ as $\mathbb{E}B_i(1 - (1 \wedge xD_i/B_i)) = 0$, which, for $x > 0$, is equivalent to $D_i/B_i \geq 1/x$ a.s. Hence $\alpha_i = \inf\{x > 0: D_i/B_i \geq 1/x \text{ a.s.}\}$ and $1/\alpha_i = \sup\{y > 0: D_i/B_i \geq y \text{ a.s.}\}$. In the right-hand side of the latter equation we see the definition of $\inf D_i/B_i$.

Now consider the case $\alpha_i = \infty$. Assume that $\inf D_i/B_i = y > 0$, then $D_i/y \geq B_i$ a.s. and $g_i(1/y) = \rho_i$. On the other hand, since $\alpha_i = \infty$, there is no $x > 0$ such that $g_i(x) = \rho_i$. Hence $y = 0$. \square

Having established the above properties of the $g_i(\cdot)$'s, we now can prove Theorem 2 by adapting a technique developed by Borst et al. [6, Lemma 5.2].

PROOF OF THEOREM 2. We first show uniqueness. Let $z^* \in (0, \infty)^J$ be an invariant NFMS, i.e., satisfy (10). Recall that $\lambda(z^*)$ is the unique optimal solution for the concave optimization problem (1). The necessary and sufficient conditions for that are given by KKT theorem (see e.g., Balder [3, Theorem 3.1]): there exist $p \in \mathbb{R}_+^J$ and $\tilde{q} \in \mathbb{R}_+^I$ such that

$$\begin{aligned} z_i^* \mathcal{U}'_i(\lambda_i(z^*)) &= z_i^* \sum_{j=1}^J A_{ji} p_j + \tilde{q}_i \quad \text{for all } i, \\ p_j \left(\sum_{i=1}^I A_{ji} \lambda_i(z^*) z_i^* - C_j \right) &= 0 \quad \text{for all } j, \\ \tilde{q}_i (\lambda_i(z^*) - m_i) &= 0 \quad \text{for all } i, \end{aligned} \tag{22}$$

or equivalently, there exist $p \in \mathbb{R}_+^J$ and $q \in \mathbb{R}_+^I$ ($q_i = \tilde{q}_i/z_i^*$) such that

$$\mathcal{U}'_i(\lambda_i(z^*)) = \sum_{j=1}^J A_{ji} p_j + q_i \quad \text{for all } i, \tag{23a}$$

$$p_j \left(\sum_{i=1}^I A_{ji} \lambda_i(z^*) - C_j \right) = 0 \quad \text{for all } j, \tag{23b}$$

$$q_i (\lambda_i(z^*) - m_i) = 0 \quad \text{for all } i. \tag{23c}$$

The theorem assumes that $m_i \leq \alpha_i$. So, by Lemmas 6 and 7, the functions $g_i(\cdot)$ are strictly increasing in the intervals $[0, m_i]$ (see also the left graph in Figure 2), which implies two things. First, the fixed point Equation (10) can be rewritten as $\lambda_i(z^*) = g_i^{-1}(\Lambda_i(z^*))$ for all i , and we plug that into (23a). Second, the second multiplier $(\lambda_i(z^*) - m_i)$ in (23c) is zero if and only if $g_i(\lambda_i(z^*)) = g_i(m_i)$, and that, by (10), is equivalent to $\Lambda_i(z^*) = g_i(m_i)$. Hence, $\Lambda(z^*)$ satisfies

$$\mathcal{U}'_i(g_i^{-1}(\Lambda_i(z^*))) = \sum_{j=1}^J A_{ji} p_j + q_i \quad \text{for all } i, \tag{24a}$$

$$p_j \left(\sum_{i=1}^I A_{ji} \Lambda_i(z^*) - C_j \right) = 0 \quad \text{for all } j, \tag{24b}$$

$$q_i (\Lambda_i(z^*) - g_i(m_i)) = 0 \quad \text{for all } i. \tag{24c}$$

Now note that the last three equations form the KKT conditions for another optimization problem. Indeed, take functions $\tilde{g}_i(\cdot)$ that are continuous and strictly increasing in \mathbb{R}_+ and coincide with $g_i(\cdot)$ in $[0, m_i]$ (and hence, the inverse functions $\tilde{g}_i^{-1}(\cdot)$ and $g_i^{-1}(\cdot)$ coincide in $[0, g_i(m_i)]$). Also take functions $G_i(\cdot)$ such that $G'_i(\cdot) = \mathcal{U}'_i(\tilde{g}_i^{-1}(\cdot))$ in $(0, \infty)$. Then (24) gives necessary and sufficient conditions for $\Lambda(z^*)$ to solve (12). Since the functions $\mathcal{U}_i(\cdot)$ are strictly concave, their derivatives $\mathcal{U}'_i(\cdot)$ are strictly decreasing. Then, since the $\tilde{g}_i^{-1}(\cdot)$'s are strictly increasing, the $G'_i(\cdot)$'s are strictly decreasing and, equivalently, the $G_i(\cdot)$'s are strictly concave, which implies that $\Lambda(z^*) = \Lambda^*$ is actually the unique solution to (12) and does not depend on z^* . Then we invert the $g_i(\cdot)$'s in the fixed point Equation (10), which implies that the fixed point $z_i^* = \Lambda_i^*/g_i^{-1}(\Lambda_i^*)$ is unique because Λ^* is unique.

The existence result follows similarly. There exists a unique optimal solution Λ^* to (12) and it satisfies the KKT conditions (24). Put $\lambda_i^* = g_i^{-1}(\Lambda_i^*)$ for all i . Then λ^* and Λ^* satisfy the KKT conditions (23) and (22), i.e., for the vector z^* with $z_i^* := \Lambda_i^*/\lambda_i^*$, we have $\lambda^* = \lambda(z^*)$ and $\Lambda^* = \Lambda(z^*)$. Plugging the last two relations into the definition of λ^* , we get the fixed point equation. \square

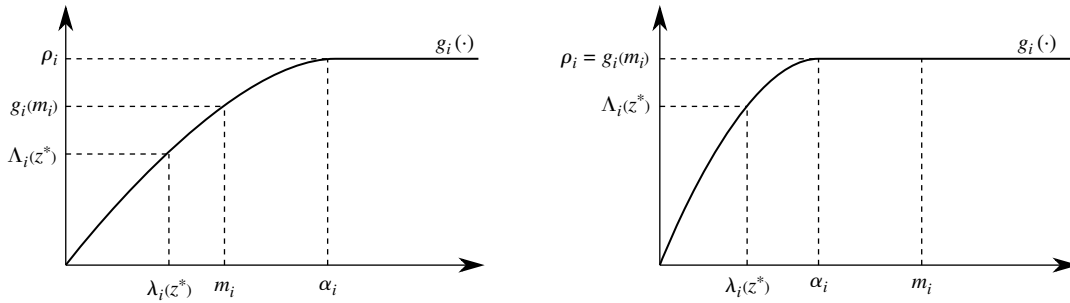


FIGURE 2. Graph of the function $g_i(\cdot)$ in the two possible cases: when $m_i \leq \alpha_i$ (left) and when $m_i > \alpha_i$ (right); z^* is an invariant NFMS.

6.3. Proof of Theorem 4. The fixed point Equation (10) and the monotonicity of the functions $g_i(\cdot)$ imply that the bandwidth class i gets in an equilibrium is at most $g_i(m_i)$. Therefore, we refer to the scenarios $\sum_{i=1}^I g_i(m_i) < C$, $\sum_{i=1}^I g_i(m_i) = C$ and $\sum_{i=1}^I g_i(m_i) > C$ as underloaded, critically loaded and overloaded, respectively. Below we calculate the invariant NFMSs in the three cases.

Summing up (9), the KKT conditions (23) for (1) and the capacity and rate constraints, a $z^* \in (0, \infty)^I$ is an invariant NFMS if and only if there exist $p \in \mathbb{R}_+$ and $q \in \mathbb{R}_+^I$ such that (we omit the argument of the rates $\lambda_i(z^*)$ and bandwidth allocations $\Lambda_i(z^*)$)

$$\Lambda_i = g_i(\lambda_i) \quad \text{for all } i, \tag{25a}$$

$$\kappa_i / \lambda_i = p + q_i \quad \text{for all } i, \tag{25b}$$

$$p \left(\sum_{i=1}^I \Lambda_i - C \right) = 0 \tag{25c}$$

$$q_i(\lambda_i - m_i) = 0 \quad \text{for all } i, \tag{25d}$$

$$\sum_{i=1}^I \Lambda_i \leq C, \tag{25e}$$

$$\lambda_i \leq m_i \quad \text{for all } i. \tag{25f}$$

Underload. In this case, there is no interaction between the classes, they do not compete but all get the maximum rate allowed. Indeed, (25c) and (25b) imply that $p = 0$ and all $q_i > 0$. Then, by (25d) and (25a), all $\lambda_i = m_i$ and all $\Lambda_i = g_i(m_i)$. Hence, there is a unique invariant NFMS given by

$$z_i^* = g_i(m_i) / m_i \quad \text{for all } i.$$

Critical load. First note that

$$\Lambda_i = g_i(m_i) \quad \text{for all } i. \tag{26}$$

Indeed, there are two possibilities: either $p = 0 \stackrel{(25b)}{\Rightarrow}$ all $q_i > 0 \stackrel{(25d)}{\Rightarrow}$ all $\lambda_i = m_i \stackrel{(25a)}{\Rightarrow}$ (26), or $p > 0 \stackrel{(25c)}{\Rightarrow} \sum_{i=1}^I \Lambda_i = C \Rightarrow$ (26), where the last implication is due to $\Lambda_i \leq g_i(m_i)$ and $\sum_{i=1}^I g_i(m_i) = C$.

Recall from Lemma 7 that

$$\alpha_i := \inf \{x: g_i(x) = \rho_i\} = 1 / \inf(D_i/B_i).$$

By (26), the relations (25a) and (25f) are equivalent to $m_i \wedge \alpha_i \leq \lambda_i \leq m_i$ (see Figure 2). Hence, (25) reduces to

$$\kappa_i / \lambda_i = p + q_i, \tag{27a}$$

$$q_i(\lambda_i - m_i) = 0, \tag{27b}$$

$$m_i \wedge \alpha_i \leq \lambda_i \leq m_i. \tag{27c}$$

Let

$$\mathcal{F}_{\text{crit}} := \{i: m_i \leq \alpha_i\}.$$

Downloaded from informas.org by [131.155.2.68] on 17 September 2014, at 02:53 . For personal use only, all rights reserved.

For $i \in \mathcal{F}_{\text{crit}}$, by (27c), we have $\lambda_i = m_i$ and $z_i = g_i(m_i)/m_i$. Then (27b) is satisfied, and (27a) implies that $p \leq \kappa_i/m_i$.

Now divide $\{1, \dots, I\} \setminus \mathcal{F}_{\text{crit}}$ into two subsets $\mathcal{F}_1 \cap \mathcal{F}_2 = \emptyset$. For $i \in \mathcal{F}_1$, put $\lambda_i = m_i$, then (as for $i \in \mathcal{F}_{\text{crit}}$) $z_i^* = g_i(m_i)/m_i$, (27b) is satisfied, and (27a) implies that $p \leq \kappa_i/m_i$. For $i \in \mathcal{F}_2$, assume $\alpha_i \leq \lambda_i < m_i$. Then $q_i = 0$ by (27b), $\kappa_i/\lambda_i = p$ by (27a), and $\kappa_i/m_i < p \leq \kappa_i/\alpha_i$. Also $z_i^* = g_i(m_i)p/\kappa_i$.

Summing up everything said above, the set of invariant NFMSs is given by

$$S_{z^*} := \bigcup_{\mathcal{F} \supseteq \mathcal{F}_{\text{crit}}} \left\{ z^*: z_i^* = g_i(m_i)/m_i \text{ for } i \in \mathcal{F} \text{ and } z_i^* = g_i(m_i)p/\kappa_i \text{ for } i \notin \mathcal{F}, \right. \\ \left. \text{where } p \in \left(\max_{i \notin \mathcal{F}} \kappa_i/m_i, \min_{i \in \mathcal{F}} \kappa_i/m_i \wedge \min_{i \notin \mathcal{F}} \kappa_i/\alpha_i \right] \right\}.$$

Equivalent descriptions of S_{z^*} are

$$S_{z^*} = \left\{ z^*: z_i^* = g_i(m_i)/m_i \text{ if } p \leq \kappa_i/m_i \text{ and } z_i^* = g_i(m_i)p/\kappa_i \text{ if } p > \kappa_i/m_i, p \in S_p \right\}, \\ S_p := \left(0, \min_{i \in \mathcal{F}_{\text{crit}}} \kappa_i/m_i \wedge \min_{i \notin \mathcal{F}_{\text{crit}}} \kappa_i/\alpha_i \right] = \left(0, \min_{1 \leq i \leq I} \kappa_i/(m_i \wedge \alpha_i) \right],$$

and

$$S_{z^*} = \{ z^*: z_i^* = g_i(m_i)/(m_i \wedge \kappa_i x), x \in S_x \}, \\ S_x := \left[\max_{1 \leq i \leq I} (m_i \wedge \alpha_i)/\kappa_i, \infty \right).$$

We now apply the last formula in a couple of simple examples.

EXAMPLE 1. If $m_i \leq \alpha_i$ for all i , then $S_x = [\max_{1 \leq i \leq I} m_i/\kappa_i, \infty)$, and $\kappa_i x \geq m_i$ for all $x \in S_x$ and all i . Hence, there is a unique invariant NFMS given by $z_i^* = g_i(m_i)/m_i$ for all i , which agrees with Theorem 2.

EXAMPLE 2. If $m_1 > \alpha_1$, $m_i \leq \alpha_i$ for $i \neq 1$ and $\alpha_1/\kappa_1 \geq \max_{i \neq 1} m_i/\kappa_i$, then, for any $\lambda_1 \in [\alpha_1, m_1]$, $z^* = (g_1(m_1)/\lambda_1, g_2(m_2)/m_2, \dots, g_I(m_I)/m_I)$ is an invariant NFMS.

Overload. In this situation, by the capacity constraint (25e), at least one class of flows does not receive the maximum service, i.e., at least one $\Lambda_i < g_i(m_i)$. We first find out which classes get the maximum service and which do not, and then calculate the unique invariant NFMS.

Who gets the maximum service. Since at least one $\Lambda_i < g_i(m_i)$, at least one $\lambda_i < m_i \wedge \alpha_i$ (see Figure 2). Then (25d), (25b), and (25c) imply that at least one $q_i = 0$, $p > 0$ and $\sum_{i=1}^I \Lambda_i = C$. At this point, we can equivalently rewrite (25) as follows: there exist $x > 0$ and $\varepsilon \in \mathbb{R}_+^I$ such that (the functions $\tilde{g}_i(\cdot)$ are introduced below)

$$\Lambda_i = g_i(\lambda_i) \Leftrightarrow \Lambda_i = \tilde{g}_i(\lambda_i), \tag{28a}$$

$$\sum_{i=1}^I g_i(\lambda_i) = C \Leftrightarrow \sum_{i=1}^I \tilde{g}_i(\lambda_i) = C \tag{28b}$$

$$\lambda_i = \kappa_i(x - \varepsilon_i), \tag{28c}$$

$$\varepsilon_i(\lambda_i - m_i) = 0, \tag{28d}$$

$$\lambda_i \leq m_i. \tag{28e}$$

For all i and $x \geq 0$, put

$$\tilde{g}_i(x) := g_i(m_i \wedge x).$$

By the rate constraints (28e), in (28), we can equivalently replace $g_i(\cdot)$ by $\tilde{g}_i(\cdot)$.

If $\lambda_i < m_i$, then, by (28d) and (28c), $\varepsilon_i = 0$ and $\lambda_i = \kappa_i x$, and hence

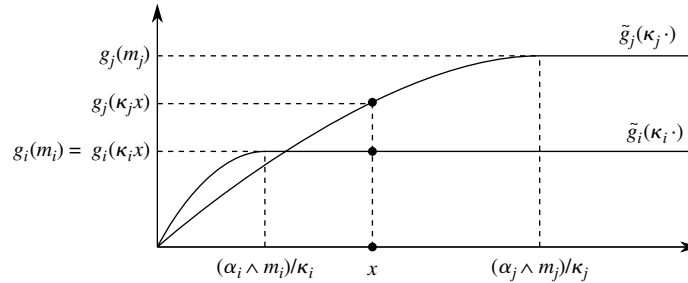
$$\tilde{g}_i(\lambda_i) = \tilde{g}_i(\kappa_i x). \tag{29}$$

If $\lambda_i = m_i$, then, by (28c), $\kappa_i x \geq m_i$ and $\tilde{g}_i(\kappa_i x) = g_i(m_i)$, and, again, (29) holds.

Plugging (29) into (28b), we get

$$\sum_{i=1}^I \tilde{g}_i(\kappa_i x) = C. \tag{30}$$

Downloaded from informas.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

FIGURE 3. Graphs of the functions $\tilde{g}_i(\cdot)$; x is the unique solution to (30).

The function $\tilde{g}(x) := \sum_{i=1}^I \tilde{g}_i(\kappa_i x)$ is continuous everywhere, strictly increasing in the interval

$$0 \leq x \leq \max_{1 \leq i \leq I} ((m_i \wedge \alpha_i) / \kappa_i) =: x_0$$

and constant for $x \geq x_0$, and also $\tilde{g}(0) = 0$ and $\tilde{g}(x_0) = \sum_{i=1}^I g_i(m_i) > C$, which implies that there exists a unique x solving (30) and $x \in (0, x_0)$.

By (28a) and (29), $\Lambda_i = \tilde{g}_i(\kappa_i x)$. Then (see Figure 3) $\Lambda_i = g_i(m_i)$ if $(m_i \wedge \alpha_i) / \kappa_i \leq x$ and $\Lambda_i < g_i(m_i)$ if $(m_i \wedge \alpha_i) / \kappa_i > x$. Hence, the set of classes that get the maximum service is

$$\mathcal{F}_{\text{over}} := \{i: (m_i \wedge \alpha_i) / \kappa_i \leq x\}. \quad (31)$$

Invariant NFMS. For $i \notin \mathcal{F}_{\text{over}}$, $\Lambda_i = g_i(\lambda_i) < g_i(m_i)$, which implies that (see Figure 2) $\lambda_i < m_i \wedge \alpha_i$. Then, by (28d) and (28c), $\varepsilon_i = 0$ and $\lambda_i = \kappa_i x$ (meeting the rate constraint (28e)), and $z_i = \Lambda_i / \lambda_i = g_i(\kappa_i x) / (\kappa_i x)$.

For $i \in \mathcal{F}_{\text{over}}$, consider the two possible cases: $\kappa_i x < m_i$ and $\kappa_i x \geq m_i$. If $\kappa_i x < m_i$, then, by (28c) and (28d), $\lambda_i \leq \kappa_i x < m_i$ and $\varepsilon_i = 0$, and, again by (28c), $\lambda_i = \kappa_i x$. If $\kappa_i x \geq m_i$, then $\lambda_i = m_i$ because otherwise we would arrive at a contradiction: $\lambda_i < m_i \Rightarrow \varepsilon_i = 0 \Rightarrow \lambda_i = \kappa_i x \geq m_i$. Hence, for $i \in \mathcal{F}_{\text{over}}$, $\lambda_i = m_i \wedge \kappa_i x$ and $z_i = \Lambda_i / \lambda_i = g_i(m_i) / (m_i \wedge \kappa_i x)$.

Summing up, the unique invariant NFMS is given by

$$z_i^* = g_i(m_i) / (m_i \wedge \kappa_i x) \quad \text{for } i \in \mathcal{F}_{\text{over}} \quad \text{and} \quad z_i^* = g_i(\kappa_i x) / (\kappa_i x) \quad \text{for } i \notin \mathcal{F}_{\text{over}},$$

where x is the unique solution to (30) and $\mathcal{F}_{\text{over}}$ is defined by (31).

7. Proof of Theorem 5. To prove C-tightness (that is, tightness with all weak limits being a.s. continuous) of $\{\tilde{\mathcal{L}}^r(\cdot)\}_{r \in \mathcal{R}}$, we check standard conditions (see, e.g., Ethier and Kurtz [12]) of compact containment (§7.2) and oscillation control (§7.4). In §7.6, we check that fluid limits satisfy the fluid model Equation (3).

To establish these main steps of the proof, we develop a number of auxiliary results. Section 7.1 contains a law of large numbers result for the load process. Section 7.3 proves that, for large r , $\tilde{\mathcal{L}}^r(\cdot)$ puts arbitrarily small mass to thin horizontal and vertical strips, which in particular implies that fluid limits have both projections free of atoms. In §7.5, fluid limits are shown to be coordinate-wise bounded away from zero outside $t = 0$.

7.1. Load process. Introduce the measure valued load processes and their scaled versions: for all r, i and $t \geq s \geq 0$,

$$\begin{aligned} \mathcal{L}_i^r(t) &:= \sum_{k=1}^{E_i^r(t)} \delta_{(B_{ik}^r, D_{ik}^r)}, & \tilde{\mathcal{L}}_i^r(t) &:= \mathcal{L}_i^r(t) / r, \\ \mathcal{L}_i^r(s, t) &:= \mathcal{L}_i^r(t) - \mathcal{L}_i^r(s), & \tilde{\mathcal{L}}_i^r(s, t) &:= \tilde{\mathcal{L}}_i^r(t) - \tilde{\mathcal{L}}_i^r(s). \end{aligned}$$

The following property is useful when proving other results of the section. Only minor adjustments in the proof of Gromoll and Williams [14, Theorem 5.1] are needed to establish it.

LEMMA 8. By (A.3) and (A.4), as $r \rightarrow \infty$, $(\tilde{\mathcal{L}}^r(\cdot), \langle \chi_1, \tilde{\mathcal{L}}^r(\cdot) \rangle, \langle \chi_2, \tilde{\mathcal{L}}^r(\cdot) \rangle) \Rightarrow (\eta(\cdot) * \theta, \rho(\cdot), \sigma(\cdot))$, where $\chi_1(x_1, x_2) := x_1$, $\chi_2(x_1, x_2) := x_2$, and $\eta(t) := t\eta$, $\rho(t) := t\rho$, $\sigma(t) := t\sigma$.

7.2. Compact containment. The property we prove here, together with the oscillation control result that follows in §7.4, implies tightness of the scaled state descriptors.

LEMMA 9. By (A.3)–(A.5), for any $T > 0$ and $\varepsilon > 0$, there exists a compact set $K \subset \mathbf{M}^I$ such that

$$\liminf_{r \rightarrow \infty} \mathbb{P}^r \{ \tilde{\mathcal{X}}^r(t) \in K \text{ for all } t \in [0, T] \} \geq 1 - \varepsilon.$$

PROOF. Fix T and ε . It suffices to show that, for each i , there exists a compact set $K_i \subset \mathbf{M}$ such that

$$\liminf_{r \rightarrow \infty} \mathbb{P}^r \{ \tilde{\mathcal{X}}_i^r(t) \in K_i \text{ for all } t \in [0, T] \} \geq 1 - \varepsilon/I. \quad (32)$$

We use the following criterion (see, e.g., Kallenberg [17, Theorem 15.7.5]).

PROPOSITION 1. A set $\mathcal{M} \subset \mathbf{M}$ is relatively compact if and only if $\sup_{\xi \in \mathcal{M}} \xi(\mathbb{R}_+^2) < \infty$ and $\sup_{\xi \in \mathcal{M}} \xi \cdot (\mathbb{R}_+^2 \setminus [0, n]^2) \rightarrow 0$ as $n \rightarrow \infty$.

Note that

$$\tilde{\mathcal{X}}_i^r(t)(\mathbb{R}_+^2) = \bar{Z}_i^r(t) \leq \bar{Z}_i^r(0) + \bar{E}_i^r(T) = \tilde{\mathcal{X}}_i^r(0)(\mathbb{R}_+^2) + \tilde{\mathcal{L}}_i^r(T)(\mathbb{R}_+^2). \quad (33)$$

Also note that, if the residual size (patience time) of a flow at time t exceeds n , then its initial size (patience time), must have exceeded n , too, which implies the following bound:

$$\tilde{\mathcal{X}}_i^r(t)(\mathbb{R}_+^2 \setminus [0, n]^2) \leq \tilde{\mathcal{X}}_i^r(0)(\mathbb{R}_+^2 \setminus [0, n]^2) + \tilde{\mathcal{L}}_i^r(T)(\mathbb{R}_+^2 \setminus [0, n]^2). \quad (34)$$

The sequence $\{\tilde{\mathcal{X}}_i^r(0) + \tilde{\mathcal{L}}_i^r(T)\}_{r \in \mathbb{R}}$ converges and hence is tight, i.e., there exists a compact set $K'_i \subset \mathbf{M}$ such that

$$\inf_{r \in \mathbb{R}} \mathbb{P}^r \{ \tilde{\mathcal{X}}_i^r(0) + \tilde{\mathcal{L}}_i^r(T) \in K'_i \} \geq 1 - \varepsilon/I. \quad (35)$$

Put

$$K''_i := \{ \xi \in \mathbf{M}: \text{for some } \xi' \in K'_i, \xi(\mathbb{R}_+^2) \leq \xi'(\mathbb{R}_+^2) \text{ and} \\ \xi(\mathbb{R}_+^2 \setminus [0, n]^2) \leq \xi'(\mathbb{R}_+^2 \setminus [0, n]^2), n \in \mathbb{N} \}.$$

Then the criterion of relative compactness for K''_i follows from that for K'_i , and (33)–(35) imply (32) with K_i taken as the closure of K''_i . \square

7.3. Asymptotic regularity. This section contains three lemmas. Lemmas 10 and 11 prove that neither initial nor newly arriving flows concentrate along horizontal and vertical lines. These two results are combined in Lemma 12 that implies the oscillation control result of the next section, and also is useful when deriving the limiting equations for the state descriptors in §7.6.

Recall from §5 that, for $b \geq a \geq 0$,

$$H_a^b = \mathbb{R}_+ \times [a, b], \quad V_a^b = [a, b] \times \mathbb{R}_+,$$

and introduce similar notations

$$H_a^\infty := \mathbb{R}_+ \times [a, \infty), \quad V_a^\infty := [a, \infty) \times \mathbb{R}_+.$$

LEMMA 10. By (A.5) and (A.6), for any $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that

$$\liminf_{r \rightarrow \infty} \mathbb{P}^r \left\{ \sup_{x \in \mathbb{R}_+} (\| \tilde{\mathcal{X}}^r(0)(H_x^{x+a}) \| \vee \| \tilde{\mathcal{X}}^r(0)(V_x^{x+a}) \|) \leq \delta \right\} \geq 1 - \varepsilon.$$

PROOF. Fix δ and ε . Since, for any $\xi \in \mathbf{M}^I$ and $a > 0$,

$$\sup_{x \in \mathbb{R}_+} (\| \xi(H_x^{x+a}) \| \vee \| \xi(V_x^{x+a}) \|) \leq 2 \sup_{n \in \mathbb{N}} (\| \xi(H_{(n-1)a}^{na}) \| \vee \| \xi(V_{(n-1)a}^{na}) \|),$$

it suffices to find an a such that

$$\liminf_{r \rightarrow \infty} \mathbb{P}^r \{ \tilde{\mathcal{X}}^r(0) \in \mathcal{M}_a \} \geq 1 - \varepsilon,$$

where $\mathcal{M}_a := \{ \xi \in \mathbf{M}^I: \sup_{n \in \mathbb{N}} (\| \xi(H_{(n-1)a}^{na}) \| \vee \| \xi(V_{(n-1)a}^{na}) \|) < \delta/2 \}$.

The set \mathcal{M}_a is open because $\xi^k \xrightarrow{w} \xi \in \mathcal{M}_a$ implies that $\xi^k \in \mathcal{M}_a$ for k large enough. Indeed, pick an $N \in \mathbb{N}$ such that $\|\xi(H_{Na}^\infty)\| \vee \|\xi(V_{Na}^\infty)\| < \delta/2$. Then, by the portmanteau theorem,

$$\begin{aligned} & \overline{\lim}_{k \rightarrow \infty} \sup_{n \in \mathbb{N}} (\|\xi^k(H_{(n-1)a}^{na})\| \vee \|\xi^k(V_{(n-1)a}^{na})\|) \\ & \leq \overline{\lim}_{k \rightarrow \infty} \max_{1 \leq n \leq N} (\|\xi^k(H_{(n-1)a}^{na})\| \vee \|\xi^k(V_{(n-1)a}^{na})\| \vee \|\xi^k(H_{Na}^\infty)\| \vee \|\xi^k(V_{Na}^\infty)\|) \\ & \leq \max_{1 \leq n \leq N} (\|\xi(H_{(n-1)a}^{na})\| \vee \|\xi(V_{(n-1)a}^{na})\| \vee \|\xi(H_{Na}^\infty)\| \vee \|\xi(V_{Na}^\infty)\|) < \delta/2. \end{aligned}$$

By (A.6) and Gromoll and Williams [14, Lemma A.1], there exists an a such that $\mathbb{P}\{\xi^0 \in \mathcal{M}_a\} \geq 1 - \varepsilon$. Then, again by the portmanteau theorem,

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \{\tilde{\mathcal{L}}^r(0) \in \mathcal{M}_a\} \geq \mathbb{P}\{\xi^0 \in \mathcal{M}_a\} \geq 1 - \varepsilon. \quad \square$$

Besides being used in the proof of the fluid limit theorem, the following result is also used when establishing convergence of the stationary distributions of the scaled state descriptors; see §5.

LEMMA 11. *By (A.3) and (A.4), for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \underbrace{\sup_{t \in [0, T]} \sup_{x \in \mathbb{R}_+} (\|\tilde{\mathcal{L}}^{r, \text{new}}(t)(H_x^{x+a})\| \vee \|\tilde{\mathcal{L}}^{r, \text{new}}(t)(V_x^{x+a})\|) \leq \delta}_{=: \Omega_r^*} \right\} \geq 1 - \varepsilon.$$

PROOF. Fix T , δ and ε . We first construct auxiliary events Ω_0^r such that $\lim_{r \rightarrow \infty} \mathbb{P}^r \{\Omega_0^r\} \geq 1 - \varepsilon$, and then show that $\Omega_r^* \supseteq \Omega_0^r$ for all r , which implies the theorem.

Definition of Ω_0^r . By Lemma 9, there exists a compact set $K \subset \mathbf{M}^l$ such that

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \underbrace{\tilde{\mathcal{L}}^r(t) \in K \text{ for all } t \in [0, T]}_{=: \Omega_0^r} \right\} \geq 1 - \varepsilon,$$

and by Proposition 1, $M := \sup_{\xi \in K} \|\xi(\mathbb{R}_+^2)\| < \infty$ and $\sup_{\xi \in K} \|\xi(\mathbb{R}_+^2) \setminus [0, L]^2\| \leq \delta/4$ for a large enough L .

For each i , the rate function $\lambda_i(\cdot)$ is positive on $\{z \in \mathbb{R}_+^l : z_i > 0\}$ and, by Lemma 1, it is continuous there. Hence,

$$\lambda_* := \min_{1 \leq i \leq l} \inf \{ \lambda_i(z) : z_i \geq \delta/4, \|z\| \leq M \} > 0. \tag{36}$$

Put

$$\gamma := \frac{\delta}{72\|\eta\|} \wedge T \quad \text{and} \quad a := \frac{\gamma(\lambda_* \wedge 1)}{3}.$$

Also pick an N large enough so that

$$Na > L + (\|m\| \vee 1)T.$$

For $n, n' \in \mathbb{N}$, define the sets

$$\begin{aligned} I_{n, n'} & := [(n-1)a, na) \times [(n'-1)a, n'a), \\ I^{n, n'} & := [(n-2)^+a, (n+1)a) \times [(n'-2)^+a, (n'+1)a), \end{aligned}$$

and pick functions $g_{n, n'} \in \mathbf{C}_{\mathbb{R}_+^2 \rightarrow [0, 1]}$ such that

$$\mathbb{1}_{I_{n, n'}}(\cdot) \leq g_{n, n'}(\cdot) \leq \mathbb{1}_{I^{n, n'}}(\cdot).$$

Since θ is a vector of probability measures,

$$\sum_{n, n' \in \mathbb{N}} \|\langle g_{n, n'}, \theta \rangle\| \leq \left\| \sum_{n, n' \in \mathbb{N}} \theta(I^{n, n'}) \right\| \leq 9. \tag{37}$$

By Lemma 8 and the continuous mapping theorem, for all $n, n' \in \mathbb{N}$, $\langle g_{n, n'}, \tilde{\mathcal{L}}^r(\cdot) \rangle \Rightarrow \eta(\cdot) \langle g_{n, n'}, \theta \rangle$ as $r \rightarrow \infty$. Since the limits are deterministic, we have convergence in probability. Since the limits are continuous, we have uniform convergence on compact sets. Hence,

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \underbrace{\max_{1 \leq n, n' \leq N} \sup_{t \in [0, T]} \|\langle g_{n, n'}, \tilde{\mathcal{L}}^r(t) \rangle - t\eta * \langle g_{n, n'}, \theta \rangle\| \leq \delta/(16N^2)}_{=: \Omega_2^*} \right\} = 1.$$

Downloaded from informas.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

Similarly, by (A.3),

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \underbrace{\sup_{t \in [0, T]} \|\bar{E}^r(t) - t\eta\|}_{=: \Omega_3^r} \leq \delta/16 \right\} = 1.$$

For all r , put

$$\Omega_0^r := \Omega_1^r \cap \Omega_2^r \cap \Omega_3^r,$$

then $\lim_{r \rightarrow \infty} \mathbb{P}^r \{\Omega_0^r\} \geq 1 - \varepsilon$, and it is left to show that $\Omega_0^r \subseteq \Omega_*^r$.

PROOF OF $\Omega_0^r \subseteq \Omega_*^r$. Fix $r \in \mathcal{R}$, $t \in [0, T]$, $x \in \mathbb{R}_+$ and i . Also fix an outcome $\omega \in \Omega_0^r$. All random objects in the rest of the proof will be evaluated at this ω . We have to check that

$$\tilde{\mathcal{X}}_i^{r, \text{new}}(t)(H_x^{x+a}) \leq \delta, \tag{38a}$$

$$\tilde{\mathcal{X}}_i^{r, \text{new}}(t)(V_x^{x+a}) \leq \delta. \tag{38b}$$

We will show (38a), (38b) follows similarly.

Define the random time $\tau := \sup\{s \leq t: \bar{Z}_i^{r, \text{new}}(s) < \delta/4\}$ (supremum over the empty set equals 0 by convention). Although in general τ is not a continuity point for $\bar{Z}_i^{r, \text{new}}(\cdot)$, we still can estimate $\bar{Z}_i^{r, \text{new}}(\tau)$:

$$\bar{Z}_i^{r, \text{new}}(\tau) \leq \delta/2. \tag{39}$$

Indeed, if $\tau = 0$, then $\bar{Z}_i^{r, \text{new}}(\tau) = 0$, and (39) holds. If $\tau > 0$, pick a $\tau' \in [(\tau - \gamma)^+, \tau]$ such that $\bar{Z}_i^{r, \text{new}}(\tau') < \delta/4$. Then, by the definition of Ω_3^r ,

$$\bar{Z}_i^{r, \text{new}}(\tau) \leq \bar{Z}_i^{r, \text{new}}(\tau') + (\bar{E}_i^r(\tau) - \bar{E}_i^r(\tau')) \leq \delta/4 + \eta_i(\tau - \tau') + \delta/8 \leq \|\eta\|\gamma + 3\delta/8,$$

and (39) holds by the choice of γ .

Now, if $\tau = t$, then (39) implies (38a), and the proof is finished. Assume that $\tau < t$. Then, by the choice of L and (39),

$$\begin{aligned} \tilde{\mathcal{X}}_i^{r, \text{new}}(t)(H_x^{x+a}) &\leq \tilde{\mathcal{X}}_i^{r, \text{new}}(t)(H_x^{x+a} \cap [0, L]^2) + \delta/4 \\ &\leq \underbrace{\bar{Z}_i^{r, \text{new}}(\tau)}_{\leq \delta/2} + \frac{1}{r} \sum_{E_i^r(\tau)+1}^{E_i^r(t)} \underbrace{\mathbb{1}_{H_x^{x+a} \cap [0, L]^2}(B_{ik}^r - S_i(\bar{Z}^r, U_{ik}^r, t), D_{ik}^r - (t - U_{ik}^r))}_{=: s_k} + \delta/4 \end{aligned}$$

and in order to have (38a), it suffices to show that

$$\Sigma := \frac{1}{r} \sum_{E_i^r(\tau)+1}^{E_i^r(t)} s_k = \sum_{n, n' \in \mathbb{N}} \underbrace{\frac{1}{r} \sum_{E_i^r(\tau)+1}^{E_i^r(t)} s_k \mathbb{1}_{I_{n, n'}}(B_{ik}^r, D_{ik}^r)}_{=: \Sigma_{n, n'}} \leq \delta/4. \tag{40}$$

First note that

$$\Sigma_{n, n'} = 0 \quad \text{if } n > N \text{ or } n' > N. \tag{41}$$

Indeed, consider a flow on route i that arrived at $U_{ik}^r \in (\tau, t]$ with $(B_{ik}^r, D_{ik}^r) \in I_{n, n'}$. If $n > N$, then $B_{ik}^r > L + \|m\|T$ by the choice of N , $B_{ik}^r - S_i(\bar{Z}^r, U_{ik}^r, t) > L$ by the rate constraints, and $s_k = 0$. If $n' > N$, then $D_{ik}^r > L + T$ by the choice of N , $D_{ik}^r - (t - U_{ik}^r) > L$ and again $s_k = 0$.

Now we estimate $\Sigma_{n, n'}$ for $1 \leq n, n' \leq N$. Fix n, n' . Consider two flows $k < l$ such that $U_{ik}^r, U_{il}^r \in (\tau, t]$ and $(B_{ik}^r, D_{ik}^r), (B_{il}^r, D_{il}^r) \in I_{n, n'}$. In $(\tau, t]$, $\bar{Z}_i^r(\cdot) \geq \bar{Z}_i^{r, \text{new}}(\cdot) \geq \delta/4$ and $\|\bar{Z}^r(\cdot)\| \leq M$, and then (36) implies that

$$\inf_{s \in (\tau, t]} \lambda_i(\bar{Z}^r(s)) \geq \lambda_*.$$

If $U_{il}^r - U_{ik}^r \geq \gamma$, then

$$\begin{aligned} (B_{il}^r - S_i(\bar{Z}^r, U_{il}^r, t)) - (B_{ik}^r - S_i(\bar{Z}^r, U_{ik}^r, t)) &\geq \underbrace{\gamma \lambda_*}_{\geq 3a} - \underbrace{(B_{ik}^r - B_{il}^r)}_{\leq a} \geq 2a, \\ (D_{il}^r - (t - U_{il}^r)) - (D_{ik}^r - (t - U_{ik}^r)) &\geq \underbrace{\gamma}_{\geq 3a} - \underbrace{(D_{ik}^r - D_{il}^r)}_{\leq a} \geq 2a, \end{aligned}$$

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

and hence at most one of s_k and s_j is nonzero. This implies that all arrivals to route i during $(\tau, t]$ that correspond to nonzero summands in $\Sigma_{n,n'}$ occur actually during a smaller interval $(t_{n,n'}, t_{n,n'} + \gamma] \subseteq (\tau, t]$. Then, by the definition of Ω_2^r ,

$$\begin{aligned} \Sigma_{n,n'} &\leq \frac{1}{r} \sum_{k=E_i^r(t_{n,n'})+1}^{E_i^r(t_{n,n'}+\gamma)} \mathbb{1}_{t_{n,n'}}(B_{ik}^r, D_{ik}^r) \leq \sup_{s \in [0, T-\gamma]} \frac{1}{r} \sum_{k=E_i^r(s)+1}^{E_i^r(s+\gamma)} g_{n,n'}(B_{ik}^r, D_{ik}^r) \\ &= \sup_{s \in [0, T-\gamma]} (\langle g_{n,n'}, \tilde{\mathcal{L}}_i^r(s+\gamma) \rangle - \langle g_{n,n'}, \tilde{\mathcal{L}}_i^r(s) \rangle) \leq \gamma \eta_i \langle g_{n,n'}, \theta_i \rangle + \delta / (8N^2). \end{aligned}$$

We plug the last inequality and (41) into $\Sigma = \sum_{n,n' \in \mathbb{N}} \Sigma_{n,n'}$, then (40) follows by (37) and the choice of γ . \square

The previous two lemmas are summed up into the following result.

LEMMA 12. *By (A.3)–(A.6), for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \sup_{t \in [0, T]} \sup_{x \in \mathbb{R}_+} \|\tilde{\mathcal{X}}^r(t)(H_x^{x+a})\| \vee \|\tilde{\mathcal{X}}^r(t)(V_x^{x+a})\| \leq \delta \right\} \geq 1 - \varepsilon.$$

PROOF. Note that

$$\begin{aligned} &\sup_{x \in \mathbb{R}_+} \|\tilde{\mathcal{X}}^{r, \text{init}}(t)(H_x^{x+a})\| \vee \|\tilde{\mathcal{X}}^{r, \text{init}}(t)(V_x^{x+a})\| \\ &\leq \sup_{x \in \mathbb{R}_+} \|\tilde{\mathcal{X}}^r(0)(H_x^{x+a})\| \vee \|\tilde{\mathcal{X}}^r(0)(V_x^{x+a})\|. \end{aligned}$$

Indeed,

$$\tilde{\mathcal{X}}_i^{r, \text{init}}(t)(H_x^{x+a}) \leq \tilde{\mathcal{X}}_i^r(0)(H_{x+t}^{x+a+t}) \quad \text{and} \quad \tilde{\mathcal{X}}_i^{r, \text{init}}(t)(V_x^{x+a}) \leq \tilde{\mathcal{X}}_i^r(0)(V_{x+S_i(\tilde{Z}^r, 0, t)}^{x+a+S_i(\tilde{Z}^r, 0, t)}).$$

Then the lemma follows by $\tilde{\mathcal{X}}^r(\cdot) = (\tilde{\mathcal{X}}^{r, \text{init}} + \tilde{\mathcal{X}}^{r, \text{new}})(\cdot)$ and Lemmas 10 and 11. \square

7.4. Oscillation control. Here we establish the second key ingredient of tightness of the scaled state descriptors, the first one is proven in §7.2.

LEMMA 13. *By (A.3)–(A.6), for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $h > 0$ such that*

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \underbrace{\omega(\tilde{\mathcal{X}}^r, h, T) \leq \delta}_{=: \Omega_4^r} \right\} \geq 1 - \varepsilon,$$

where $\omega(\tilde{\mathcal{X}}^r, h, T) := \sup\{d_T(\tilde{\mathcal{X}}^r(s), \tilde{\mathcal{X}}^r(t)) : s, t \in [0, T], |s - t| < h\}$.

PROOF. Fix T , δ and ε . By (A.3),

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \underbrace{\sup_{t \in [0, T]} \|\bar{E}^r(t) - t\eta\| \leq \delta/4}_{=: \Omega_1^r} \right\} = 1.$$

By Lemma 12, there exists an $a > 0$ such that

$$\lim_{r \rightarrow \infty} \mathbb{P}^r \left\{ \underbrace{\sup_{t \in [0, T]} \|\tilde{\mathcal{X}}^r(t)(H_0^a \cup V_0^a)\| \leq \delta}_{=: \Omega_2^r} \right\} \geq 1 - \varepsilon.$$

Pick an h such that $h(\|m\| \vee 1) \leq \delta \wedge a$ and $h\|\eta\| \leq \delta/2$. We now show that, for all $r \in \mathcal{R}$, $\Omega_* \supseteq \Omega_1^r \cap \Omega_2^r$, then the lemma follows.

Fix $r \in \mathcal{R}$, i and $s, t \in [0, T]$ such that $s < t$, $t - s < h$. Also fix an outcome $\omega \in \Omega_1^r \cap \Omega_2^r$. All random objects in the rest of the proof will be evaluated at this ω . We have to check that, for any nonempty closed Borel subset $B \subseteq \mathbb{R}_+^2$,

$$\tilde{\mathcal{X}}_i^r(s)(B) \leq \tilde{\mathcal{X}}_i^r(t)(B^\delta) + \delta, \tag{42a}$$

$$\tilde{\mathcal{X}}_i^r(t)(B) \leq \tilde{\mathcal{X}}_i^r(s)(B^\delta) + \delta. \tag{42b}$$

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

First we check (42a). Note that it suffices to show

$$\tilde{\mathcal{X}}_i^r(s)(B) \leq \tilde{\mathcal{X}}_i^r(\tau)(B^\delta) + \delta, \tag{43}$$

where $\tau := \inf\{u \in [s, t]: \tilde{Z}_i^r(u) = 0\}$ and infimum over the empty set equals t by definition. Indeed, if $\tau = t$, then (43) implies (42a). If $\tau < t$, then by the right continuity of $\tilde{Z}_i^r(\cdot)$, $\tilde{\mathcal{X}}_i^r(\tau)(B^\delta) = \tilde{Z}_i^r(\tau) = 0$, and again (43) implies (42a).

Now prove (43). If $\tau = s$, then (43) holds. Assume that $\tau > s$. By the definition of Ω_2^r ,

$$\tilde{\mathcal{X}}_i^r(s)(B) \leq \tilde{\mathcal{X}}_i^r(s)(B \cap [a, \infty)^2) + \delta. \tag{44}$$

Since $S_i(\tilde{Z}^r, s, \tau) < \|m\|h \leq \delta \wedge a$ and $\tau - s < h \leq \delta \wedge a$,

$$\tilde{\mathcal{X}}_i^r(s)(B \cap [a, \infty)^2) \leq \tilde{\mathcal{X}}_i^r(\tau)(B^\delta),$$

which together with (44) implies (43).

It is left to check (42b). Since $S_i(\tilde{Z}^r, s, \tau) < \|m\|h \leq \delta$ and $\tau - s < h \leq \delta$,

$$\tilde{\mathcal{X}}_i^r(t)(B) \leq \tilde{\mathcal{X}}_i^r(s)(B^\delta) + (\bar{E}_i^r(t) - \bar{E}_i^r(s)),$$

and (42b) follows by the definition of Ω_1^r . \square

7.5. Fluid limits are bounded away from zero. Rate constraints provide infinite-server-queue lower bounds for bandwidth-sharing networks. First we show that properly scaled infinite server queues are bounded away from zero, and then the same follows for bandwidth-sharing networks with rate constraints.

Consider a sequence of infinite server queues marked by $r \in \mathcal{R}$. At $t = 0$, the queues are empty. To the r th queue, customers arrive according to a counting process $A^r(\cdot)$ and have i.i.d. service times $\{B_k^r\}_{k \in \mathbb{N}}$ distributed as B^r . Let $\bar{A}^r(\cdot) := A^r(\cdot)/r \Rightarrow \alpha(\cdot)$, where $\alpha(t) := t\alpha$ and $\alpha > 0$. Also let $B^r \Rightarrow B$, where $\mathbb{P}\{B > 0\} > 0$. Denote by $Q^r(\cdot)$ the population process of the r th queue and put $\bar{Q}^r(\cdot) := Q^r(\cdot)/r$.

LEMMA 14. *For any $\delta > 0$, there exists a $C(\delta) > 0$ such that, for any $\Delta > \delta$,*

$$\mathbb{P}^r \left\{ \inf_{\delta \leq t \leq \Delta} \bar{Q}^r(t) \geq C(\delta) \right\} \rightarrow 1 \quad \text{as } r \rightarrow \infty.$$

PROOF. Let us first explain the result heuristically. Consider the arrivals with long service times, i.e., exceeding a $b > 0$. During $(0, b/2]$, there are $r\alpha\mathbb{P}\{B > b\}b/2$ such arrivals to the r th queue. They will leave the queue after $t = b$, and hence, in $(b/2, b]$, the scaled queue length $\bar{Q}^r(\cdot)$ is bounded from below by $\alpha\mathbb{P}\{B > b\}b/2$. Similarly, $\bar{Q}^r(\cdot) \geq \alpha\mathbb{P}\{B > b\}b/2$ in any interval $((n-1)b/2, nb/2]$, $n \in \mathbb{N}$.

We now proceed more formally. Pick an $b \in (0, \delta)$ such that b is a continuity point for the distribution of B , and

$$p := \mathbb{P}\{B \geq b\} > 0.$$

Then, as $r \rightarrow \infty$,

$$p_r := \mathbb{P}^r\{B^r \geq b\} \rightarrow p.$$

Partition $(0, \Delta]$ into subintervals of length $b/2$,

$$(0, \Delta] \subseteq \bigcup_{1 \leq n \leq N(\Delta)} ((n-1)b/2, nb/2].$$

Denote by \bar{A}_n^r the scaled number of arrivals during $((n-1)b/2, nb/2]$, and by $\bar{A}_n^r(b)$ the scaled number of arrivals during $((n-1)b/2, nb/2]$ with service times at least b ,

$$\begin{aligned} \bar{A}_n^r &:= \bar{A}^r(nb/2) - \bar{A}^r((n-1)b/2), \\ \bar{A}_n^r(b) &:= \frac{1}{r} \sum_{k=A^r((n-1)b/2)+1}^{A^r(nb/2)} \mathbb{1}_{\{B_k^r \geq b\}}. \end{aligned}$$

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

By $\bar{A}^r \Rightarrow \alpha(\cdot)$ and $p_r \rightarrow p$ as $r \rightarrow \infty$,

$$\begin{aligned} (\bar{A}_1^r, \dots, \bar{A}_{N(\Delta)}^r) &\Rightarrow (\alpha b/2, \dots, \alpha b/2), \\ (\bar{A}_1^r(b), \dots, \bar{A}_{N(\Delta)}^r(b)) &\Rightarrow (\alpha p b/2, \dots, \alpha p b/2). \end{aligned}$$

Pick a $C(\delta) < \alpha p b/2$, then

$$\begin{aligned} \mathbb{P}^r \left\{ \inf_{\delta \leq t \leq \Delta} \bar{Q}^r(t) \geq C(\delta) \right\} &\geq \mathbb{P}^r \left\{ \inf_{t \in ((n-1)b/2, nb/2]} \bar{Q}^r(t) \geq C(\delta), n = 2, \dots, N(\Delta) \right\} \\ &\geq \mathbb{P}^r \left\{ \bar{A}_n^r(b) \geq C(\delta), n = 1, \dots, N(\Delta) - 1 \right\} \rightarrow 1 \quad \text{as } r \rightarrow \infty. \quad \square \end{aligned}$$

We can now prove easily that all fluid limits are bounded away from zero outside $t = 0$.

LEMMA 15. *For any $\delta > 0$, there exists a $C(\delta) > 0$ such that, for any fluid limit $(\mathcal{X}, Z)(\cdot)$,*

$$\text{a.s.} \quad \inf_{t \geq \delta} \min_{1 \leq i \leq I} Z_i(t) \geq C(\delta).$$

PROOF. Consider a flow k on route i in the r th network. By the rate constraints, this flow will stay in the network at least for $B_{ik}^r/m_i \wedge D_{ik}^r$ since its arrival. Hence, the route i population process $Z_i^r(\cdot)$ is bounded from below by the length $Q_i^r(\cdot)$ of the infinite server queue with arrivals $E_i^r(\cdot)$ and i.i.d. service times $\{B_{ik}^r/m_i \wedge D_{ik}^r\}_{k \in \mathbb{N}}$. Assume that $Q_i^r(0) = 0$ and put $\bar{Q}_i^r(\cdot) = Q_i^r(\cdot)/r$. Then, by Lemma 14, for any $\delta > 0$ there exists a $C(\delta) > 0$ such that, for any $\Delta > \delta$,

$$\mathbb{P}^r \left\{ \inf_{t \in [\delta, \Delta]} \min_{1 \leq i \leq I} \bar{Z}_i^r(t) \geq C(\delta) \right\} \geq \mathbb{P}^r \left\{ \inf_{t \in [\delta, \Delta]} \min_{1 \leq i \leq I} \bar{Q}_i^r(t) \geq C(\delta) \right\} \rightarrow 1.$$

Now consider a fluid limit $(\mathcal{X}, Z)(\cdot)$ along a subsequence $\{(\bar{\mathcal{X}}^q, \bar{Z}^q)(\cdot)\}_{q \in \mathcal{Q}}$. For any compact set $K \subset \mathbb{R}_+$, the mapping $\varphi_K: \mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbb{R}} \rightarrow \mathbb{R}$, $\varphi_K(x) := \inf_{t \in K} \min_{q \leq i \leq I} x(t)$ is continuous at continuous $x(\cdot)$. Hence, $\varphi_{[\delta, \Delta]}(\bar{Z}^q) \Rightarrow \varphi_{[\delta, \Delta]}(Z)$ and, by the portmanteau theorem,

$$\mathbb{P}\{\varphi_{[\delta, \Delta]}(Z) \geq C(\delta)\} \geq \overline{\lim}_{q \rightarrow \infty} \mathbb{P}^q\{\varphi_{[\delta, \Delta]}(\bar{Z}^q) \geq C(\delta)\} = 1,$$

where $\Delta > \delta$ is arbitrary. Then the lemma follows.

Note also that the constant $C(\delta)$ does not depend on a particular fluid limit $(\mathcal{X}, Z)(\cdot)$. \square

7.6. Fluid limits as fluid model solutions. Here we show that fluid limits a.s. satisfy the fluid model Equation (3).

Let $(\mathcal{X}, Z)(\cdot)$ be a fluid limit along a subsequence $\{(\bar{\mathcal{X}}^q, \bar{Z}^q)(\cdot)\}_{q \in \mathcal{Q}}$. Lemma 12 implies that (cf. the proof of Gromoll et al. [15, Lemma 6.2])

$$\text{a.s.} \quad \mathcal{X}_i(t)(\partial_A) = 0 \quad \text{for all } t \geq 0, \text{ all } i \text{ and } A \in \mathcal{C}, \tag{45}$$

where ∂_A denotes the boundary of A . Then, when proving (3) for $(\mathcal{X}, Z)(\cdot)$, it suffices to consider sets A from

$$\mathcal{C}^+ := \{[x, \infty) \times [y, \infty): x \wedge y > 0\}.$$

It also suffices to consider t from a finite interval $[0, T]$.

The rest of the proof splits into two parts. First we derive dynamic equations for the prelimiting processes $(\bar{\mathcal{X}}^q, \bar{Z}^q)(\cdot)$, and then show that these equations converge to (3).

Prelimiting equations. Fix $q \in \mathcal{Q}$ and $t \leq T$. Also fix a coordinate i and a set $A \in \mathcal{C}^+$. What follows up to and including Equation (48), holds for all possible outcomes $w \in \Omega^q$ of the probability space $(\Omega^q, \mathcal{F}^q, \mathbb{P}^q)$ on which system q is defined.

We have

$$\bar{\mathcal{X}}_i^q(t)(A) = \bar{\mathcal{X}}_i^q(0)(A + (S_i(\bar{Z}^q, 0, t), t)) + \underbrace{\frac{1}{q} \sum_{k=1}^{E_i^q(t)} \mathbb{1}_A(B_{ik}^q - S_i(\bar{Z}^q, U_{ik}^q, t), D_{ik}^q - (t - U_{ik}^q))}_{=:\Sigma} \tag{46}$$

Downloaded from informs.org by [131.155.2.68] on 17 September 2014, at 02:53. For personal use only, all rights reserved.

Fix a partition $0 < t_0 < t_1 < \dots < t_N = t$, then

$$\Sigma = \frac{1}{q} \sum_{k=1}^{E_i^q(t_0)} s_k + \frac{1}{q} \sum_{j=0}^{N-1} \sum_{k=E_i^q(t_j)+1}^{E_i^q(t_{j+1})} s_k.$$

Suppose that a function $y(\cdot)$ is nonincreasing in $[t_0, t]$ and that, for some δ ,

$$\sup_{s \in [t_0, t]} |S_i(\bar{Z}^q, s, t) - y(s)| \leq \delta.$$

Now we can estimate Σ . If $U_{ik}^q \in (t_j, t_{j+1}]$, then

$$B_{ik}^q - (y(t_j) + \delta) \leq B_{ik}^q - S(\bar{Z}^q, U_{ik}^q, t) \leq B_{ik}^q - (y(t_{j+1}) - \delta),$$

$$D_{ik}^q - (t - t_j) \leq D_{ik}^q - (t - U_{ik}^q) \leq D_{ik}^q - (t - t_{j+1}),$$

and

$$\Sigma \geq \sum_{j=0}^{N-1} \frac{1}{q} \sum_{k=E_i^q(t_j)+1}^{E_i^q(t_{j+1})} \mathbb{1}_A(B_{ik}^q - (y(t_j) + \delta), D_{ik}^q - (t - t_j)),$$

$$\Sigma \leq \bar{E}_i^q(t_0) + \sum_{j=0}^{N-1} \frac{1}{q} \sum_{k=E_i^q(t_j)+1}^{E_i^q(t_{j+1})} \mathbb{1}_A(B_{ik}^q - (y(t_{j+1}) - \delta), D_{ik}^q - (t - t_{j+1})),$$

which can be rewritten as

$$\Sigma \geq \sum_{j=0}^{N-1} \bar{\mathcal{L}}_i^q(t_j, t_{j+1})(A + (y(t_j) + \delta, t - t_j))$$

$$\Sigma \leq \bar{E}_i^q(t_0) + \sum_{j=0}^{N-1} \bar{\mathcal{L}}_i^q(t_j, t_{j+1})(A + (y(t_{j+1}) - \delta, t - t_{j+1})).$$
(47)

Put

$$X^q := \sup_{A \in \mathcal{C}} \sup_{0 \leq s \leq t \leq T} \left\| (\bar{\mathcal{L}}^q(s, t)(A) - (t - s)\eta * \theta^q(A)) \right\|,$$

then, by (47) and (46),

$$\sum_{j=0}^{N-1} (\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_j) + \delta, t - t_j)) - X^q)$$

$$\leq \bar{\mathcal{L}}_i^q(t)(A) - \bar{\mathcal{L}}_i^q(0)(A + (S_i(\bar{Z}^q, 0, t), t))$$

$$\leq \eta_i t_0 + X^q + \sum_{j=0}^{N-1} (\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_{j+1}) - \delta, t - t_{j+1})) + X^q).$$
(48)

To summarize, we have shown that, for all $q \in \mathbb{Q}$ and all possible outcomes $\omega \in \Omega^q$,

$$(\bar{\mathcal{L}}^q(\cdot), X^q) \in \mathcal{A}^q, \tag{49}$$

where $\mathcal{A}^q \subset \mathbf{D}_{\mathbb{R}_+ \rightarrow \mathbf{M}'} \times \mathbb{R}_+$ is the set of pairs $(\zeta(\cdot), x)$ such that, for any set $A \in \mathcal{C}^+$, any partition $0 < t_0 < t_1 < \dots < t_N = t \leq T$ and any function $y(\cdot)$ that is nonincreasing in $[t_0, t]$ and that satisfies $\sup_{s \in [t_0, t]} |S_i(\langle 1, \zeta \rangle, s, t) - y(s)| \leq \delta$ for some i and δ ,

$$\sum_{j=0}^{N-1} (\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_j) + \delta, t - t_j)) - x)$$

$$\leq \zeta(t)(A) - \zeta_i(0)(A + (S_i(\langle 1, \zeta \rangle, 0, t), t))$$

$$\leq \eta_i t_0 + x + \sum_{j=0}^{N-1} (\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_{j+1}) - \delta, t - t_{j+1})) + x).$$

Limiting equations. By (A.3) and (A.4) (cf. the proof of Gromoll et al. [15, Lemma 5.1]),

$$X_q \Rightarrow 0 \quad \text{as } q \rightarrow \infty.$$

Since the limit of X_q is deterministic, then the joint convergence $(\tilde{\mathcal{X}}^q(\cdot), X^q) \Rightarrow (\mathcal{X}(\cdot), 0)$ holds. By the Skorokhod representation theorem, there exist random elements $\{\tilde{\mathcal{X}}^q(\cdot)\}_{q \in \mathcal{Q}}$, $\tilde{\mathcal{X}}(\cdot)$, and $\{\tilde{X}^q\}_{q \in \mathcal{Q}}$ defined on a common probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that $(\tilde{\mathcal{X}}^q(\cdot), \tilde{X}^q) \stackrel{d}{=} (\tilde{\mathcal{X}}^q(\cdot), X^q)$, $q \in \mathcal{Q}$, and $\tilde{\mathcal{X}}(\cdot) \stackrel{d}{=} \mathcal{X}(\cdot)$, and

$$\text{a.s. } (\tilde{\mathcal{X}}^q(\cdot), \tilde{X}^q) \rightarrow (\tilde{\mathcal{X}}(\cdot), 0) \quad \text{as } q \rightarrow \infty. \quad (50)$$

Introduce also the total mass processes $\tilde{Z}^q(\cdot) := \langle 1, \tilde{\mathcal{X}}^q(\cdot) \rangle$, $q \in \mathcal{Q}$, and $\tilde{Z}(\cdot) := \langle 1, \tilde{\mathcal{X}}(\cdot) \rangle$. By Lemma 15, (45) and (49),

$$\text{a.s. } \tilde{Z}_i(t) > 0 \quad \text{for all } t > 0 \text{ and all } i, \quad (51a)$$

$$\text{a.s. } \tilde{\mathcal{X}}_i(t)(\partial_A) = 0 \quad \text{for all } t \geq 0, \text{ all } i \text{ and } A \in \mathcal{C}, \quad (51b)$$

$$\text{a.s. } (\tilde{\mathcal{X}}^q(\cdot), \tilde{X}^q) \in \mathcal{S}^q \quad \text{for all } q \in \mathcal{Q}. \quad (51c)$$

Denote by $\tilde{\Omega}_*$ the set of outcomes $\omega \in \tilde{\Omega}$ for which (50) and (51) hold. We will show that, for all $\omega \in \tilde{\Omega}_*$, all i , $t \in [0, T]$ and $A \in \mathcal{C}^+$,

$$\tilde{\mathcal{X}}_i(t)(A) = \tilde{\mathcal{X}}_i(0)(A + (S_i(\tilde{Z}, 0, t), t)) + \eta_i \int_0^t \theta_i(A + (S_i(\tilde{Z}, s, t), t - s)) ds, \quad (52)$$

and that will complete the proof of Theorem 5.

Fix $t \in [0, T]$, i and $A \in \mathcal{C}^+$. Also fix an outcome $\omega \in \tilde{\Omega}_*$. All random elements in the rest of the proof are evaluated at this ω .

By (50) and (51b),

$$\tilde{\mathcal{X}}_i^q(t)(A) \rightarrow \tilde{\mathcal{X}}_i(t)(A) \quad \text{as } q \rightarrow \infty. \quad (53)$$

By (51a), the rate constraints and the dominated convergence theorem,

$$S_i(\tilde{Z}^q, s, t) \rightarrow S_i(\tilde{Z}, s, t) \quad \text{for all } s \in [0, t] \text{ as } q \rightarrow \infty, \quad (54)$$

which in particular implies that

$$\tilde{\mathcal{X}}_i^q(0)(A + (S_i(\tilde{Z}^q, 0, t), t)) \rightarrow \tilde{\mathcal{X}}_i(0)(A + (S_i(\tilde{Z}, 0, t), t)) \quad \text{as } q \rightarrow \infty. \quad (55)$$

Fix $t_0 \in (0, t)$ and $\delta > 0$. By (51a), the function $S_i(\tilde{Z}, \cdot, t)$ is continuous in $[t_0, t]$, and the functions $S_i(\tilde{Z}^q, \cdot, t)$ are monotone. Then the point-wise convergence (54) implies uniform convergence in $[t_0, t]$, and for q large enough,

$$\sup_{s \in [t_0, t]} |S_i(\tilde{Z}^q, s, t) - S_i(\tilde{Z}, s, t)| \leq \delta. \quad (56)$$

Now fix a partition $t_0 < t_1 < \dots < t_N = t$. The bound (56) and (51c) imply that (in the definition of \mathcal{S}^q we take $y(\cdot) = S_i(\tilde{Z}, \cdot, t)$)

$$\begin{aligned} & \sum_{j=0}^{N-1} (\eta_i(t_{j+1} - t_j) \theta_i^q(A + (S_i(\tilde{Z}, t_j, t) + \delta, t - t_j)) - \tilde{X}^q) \\ & \leq \tilde{\mathcal{X}}_i^q(t)(A) - \tilde{\mathcal{X}}_i^q(A + (S_i(\tilde{Z}^q, 0, t), t)) \\ & \leq \eta_i t_0 + \tilde{X}^q + \sum_{j=0}^{N-1} (\eta_i(t_{j+1} - t_j) \theta_i^q(A + (S_i(\tilde{Z}, t_{j+1}, t) - \delta, t - t_{j+1})) + \tilde{X}^q). \end{aligned} \quad (57)$$

Since $\theta_i(\cdot \times \mathbb{R}_+)$ and $\theta_i(\mathbb{R}_+ \times \cdot)$ are probability measures, the set of $B \in \mathcal{C}$ for which $\theta_i(\partial_B) > 0$ is at most countable. By (51), $S_i(\tilde{Z}, \cdot, t)$ is strictly monotone in $[t_0, t]$. Hence, the set \mathcal{D} of $s \in [t_0, t]$ for which $\theta_i(\partial_{A+(S_i(\tilde{Z}, s, t)+\delta, t-s)}) > 0$ or $\theta_i(\partial_{A+(S_i(\tilde{Z}, s, t)-\delta, t-s)}) > 0$ is at most countable, too. In (57), let $q \rightarrow \infty$ assuming that the partition contains no points from \mathcal{D} . Then, by (50), (53), and (55),

$$\begin{aligned} & \sum_{j=0}^{N-1} \eta_i(t_{j+1} - t_j) \theta_i(A + (S_i(\tilde{Z}, t_j, t) + \delta, t - t_j)) \\ & \leq \tilde{\mathcal{X}}_i(t)(A) - \tilde{\mathcal{X}}_i(0)(A + (S_i(\tilde{Z}, 0, t), t)) \\ & \leq \eta_i t_0 + \sum_{j=0}^{N-1} \eta_i(t_{j+1} - t_j) \theta_i(A + (S_i(\tilde{Z}, t_{j+1}, t) - \delta, t - t_{j+1})). \end{aligned} \quad (58)$$

Now, in (58), let the diameter of the partition go to 0 keeping t_0 fixed. Then

$$\begin{aligned} & \eta_i \int_{t_0}^t \theta_i(A + (S_i(\tilde{Z}, s, t) + \delta, t - s)) ds \\ & \leq \tilde{\mathcal{L}}_i(t)(A) - \tilde{\mathcal{L}}_i(0)(A + (S_i(\tilde{Z}, 0, t), t)) \\ & \leq \eta_i t_0 + \eta_i \int_{t_0}^t \theta_i(A + (S_i(\tilde{Z}, s, t) - \delta, t - s)) ds. \end{aligned}$$

Finally, in the last inequality, let $\delta \rightarrow 0$ (recall (51b)) and $t_0 \rightarrow 0$, then (52) follows.

8. Proof of Theorem 6. By the discussion following Theorem 6 and Lemma 2, it is left to show tightness of the scaled stationary distributions. It suffices to show coordinate-wise tightness, so fix i . By Jakubowski [16, Theorem 2.1] and Kallenberg [17, Theorem 15.7.5], the sequence $\{\tilde{\mathcal{Y}}_i^r, \tilde{Y}_i^r\}_{r \in \mathcal{R}}$ is tight if

$$\sup_{r \in \mathcal{R}} \mathbb{E}^r \tilde{Y}_i^r < \infty, \tag{59a}$$

$$\limsup_{n \rightarrow \infty} \sup_{r \in \mathcal{R}} \mathbb{E}^r \tilde{\mathcal{Y}}_i^r(V_n^\infty) = 0, \tag{59b}$$

$$\limsup_{n \rightarrow \infty} \sup_{r \in \mathcal{R}} \mathbb{E}^r \tilde{\mathcal{Y}}_i^r(H_n^\infty) = 0, \tag{59c}$$

where $V_n^\infty = [n, \infty) \times \mathbb{R}_+$ and $H_n^\infty = \mathbb{R}_+ \times [n, \infty)$.

First check (59a). For each r , the route i population process $Z_i^r(\cdot)$ is bounded from above by the length $Q_i^r(\cdot)$ of the $M/G/\infty$ queue with the following parameters:

(Q.1) At $t = 0$, there are $Z_i^r(0)$ customers whose service times are patience times of the initial flows on route i of the r th network.

(Q.2) The input process is the route i input process of the r th network.

(Q.3) Service times of newly arriving customers are patience times of newly arriving flows on route i of the r th network.

For all r and t , $Z_i^r(t) \leq Q_i^r(t)$. As $t \rightarrow \infty$, $Z_i^r(t) \Rightarrow Y_i^r$ and $Q_i^r(t) \Rightarrow \Pi(\eta_i^r \mathbb{E}^r D_i^r)$. Hence, $Y_i^r \leq_{\text{st}} \Pi(\eta_i^r \mathbb{E}^r D_i^r)$ and $\mathbb{E}^r \tilde{Y}_i^r \leq \eta_i^r \mathbb{E}^r D_i^r / r \rightarrow \eta_i \mathbb{E} D_i$ as $r \rightarrow \infty$, which implies (59a).

Now check (59b). Note that, if at some point the residual flow size is at least n , then the initial flow size was at least n , too. Hence, $\tilde{\mathcal{L}}_i^r(\cdot)(V_n^\infty)$ is bounded from above by the length $Q_i^{r,n}(\cdot)$ of the $M/G/\infty$ queue whose initial state is as in (Q.1), newly arriving customers are newly arriving flows on route i of the r th network with initial sizes at least n , and service times of newly arriving customers are patience times of the corresponding flows. In particular, the input process for this queue is Poisson with intensity $\eta_i^r \mathbb{P}^r\{B_i^r \geq n\}$ and, by assumption (A.4), it does not depend on service times.

Let $f_n(\cdot)$ be a continuous function on \mathbb{R}_+^2 such that

$$\mathbb{1}_{V_{n+1}^\infty}(\cdot) \leq f_n(\cdot) \leq \mathbb{1}_{V_n^\infty}(\cdot).$$

Then, for all r and t ,

$$\langle f_n, \tilde{\mathcal{L}}_i^r(t) \rangle \leq \tilde{\mathcal{L}}_i^r(t)(V_n^\infty) \leq Q_i^{r,n}(t)$$

Letting $t \rightarrow \infty$, we obtain

$$\begin{aligned} \mathcal{Y}_i^r(V_{n+1}^\infty) & \leq \langle f_n, \mathcal{Y}_i^r \rangle \leq_{\text{st}} \Pi(\eta_i^r \mathbb{P}^r\{B_i^r \geq n\} \mathbb{E}^r D_i^r), \\ \mathbb{E}^r \tilde{\mathcal{Y}}_i^r(V_{n+1}^\infty) & \leq \eta_i^r \mathbb{P}^r\{B_i^r \geq n\} \mathbb{E}^r D_i^r / r, \end{aligned}$$

and then (59b) follows.

Finally, (59c) is valid because of the following lemma.

LEMMA 16. For any $r \in \mathcal{R}$, i and Borel set $S \subseteq \mathbb{R}_+$,

$$\mathcal{Y}_i^r(\mathbb{R}_+ \times S) \leq_{\text{st}} \Pi(\eta_i^r \mathbb{E}^r D_i^r \mathbb{P}^r\{\tilde{D}_i^r \in S\}),$$

where \tilde{D}_i^r has density $\mathbb{P}^r\{D_i^r > x\} / \mathbb{E}^r D_i^r$, $x \geq 0$.

PROOF. Fix $r \in \mathcal{R}$, i and a Borel set $S \subseteq \mathbb{R}_+$. It suffices to show that, for any $\delta > 0$,

$$\mathcal{Y}_i^r(\mathbb{R}_+ \times S) \leq_{\text{st}} \Pi(\eta_i^r \mathbb{E}^r D_i^r \mathbb{P}^r\{\tilde{D}_i^r \in S^\delta\}),$$

so fix $\delta > 0$.

Consider the upper bound queue $Q_i^r(\cdot)$ with parameters (Q.1)–(Q.3). Denote by $Q_i^r(t)(S^\delta)$ the number of customers in this queue whose residual service times at time t are in S^δ . Then

$$\mathcal{X}_i^r(\cdot)(\mathbb{R}_+ \times S^\delta) \leq Q_i^r(\cdot)(S^\delta).$$

Given at time t there are k customers in the queue, denote by $D_1(t), \dots, D_k(t)$ their residual service times. By Takács [28, Chap. 3.2, Theorem 2],

$$\lim_{t \rightarrow \infty} \mathbb{P}^r \{D_1(t) \leq x_1, \dots, D_k(t) \leq x_k \mid Q_i^r(t) = k\} = \mathbb{P}^r \{\tilde{D}_i^r \leq x_1\} \cdots \mathbb{P}^r \{\tilde{D}_i^r \leq x_k\},$$

which together with $Q_i^r(t) \Rightarrow \Pi(\eta_i^r \mathbb{E}^r D_i^r)$ as $t \rightarrow \infty$ implies that

$$Q_i^r(t)(S^\delta) \Rightarrow \Pi(\eta_i^r \mathbb{E}^r D_i^r \mathbb{P}^r \{\tilde{D}_i^r \in S^\delta\}).$$

Let g_δ be a continuous function on \mathbb{R}_+^2 such that

$$\mathbb{1}_{\mathbb{R}_+ \times S}(\cdot) \leq g_\delta(\cdot) \leq \mathbb{1}_{\mathbb{R}_+ \times S^\delta}(\cdot).$$

Then, for any t ,

$$\langle g^\delta, \mathcal{X}_i^r(t) \rangle \leq \mathcal{X}_i^r(t)(\mathbb{R}_+ \times S^\delta) \leq Q_i^r(t)(S^\delta),$$

and as $t \rightarrow \infty$,

$$\mathcal{Y}_i^r(\mathbb{R}_+ \times S) \leq \langle g^\delta, \mathcal{Y}_i^r \rangle \leq_{\text{st}} \Pi(\eta_i^r \mathbb{E}^r D_i^r \mathbb{P}^r \{\tilde{D}_i^r \in S^\delta\}). \quad \square$$

Acknowledgments. The research of the first and third authors is supported by the Netherlands Organization for Scientific Research (NWO) [Vidi Grant 639.032.815].

Appendix

PROOF OF LEMMA 1. It suffices to show that, for a vector $z \in \mathbb{R}_+^I$ with the first $I' < I$ coordinates positive and the rest of them zero, and a sequence $\{z^k\}_{k \in \mathbb{N}} \subset (0, \infty)^I$ such that $z^k \rightarrow z$, we have $\Lambda(z^k) \rightarrow \Lambda(z)$.

Suppose that $z^k \rightarrow z$ but $\Lambda(z^k) \not\rightarrow \Lambda(z)$. Since $\{\Lambda(z^k)\}_{k \in \mathbb{N}}$ is a subset of the compact set $\{\Lambda \in \mathbb{R}_+^I : \|\Lambda\| \leq \|C\|\}$, without loss of generality we may assume that $\Lambda(z^k) \rightarrow \Lambda' \neq \Lambda(z)$.

Recall that $\Lambda(z)$ is the unique optimal solution to

$$\text{maximize } \sum_{i=1}^I z_i \mathcal{U}_i(\Lambda_i/z_i) \quad \text{subject to } A\Lambda \leq C, \quad \Lambda \leq m * z, \tag{60}$$

where, by convention, $\Lambda_i/0 = 0$ and $0 \times (-\infty) = 0$.

For all k , $A\Lambda(z^k) \leq C$ and $\Lambda(z^k) \leq m * z^k$. Hence, Λ' is feasible for (60) and $\Lambda'_i = 0 = \Lambda_i(z)$ for $i > I'$. Since $\Lambda' \neq \Lambda(z)$ is not optimal for (60),

$$l := \sum_{i=1}^{I'} z_i \mathcal{U}_i(\Lambda_i(z)/z_i) > \sum_{i=1}^{I'} z_i \mathcal{U}_i(\Lambda'_i/z_i) =: r. \tag{61}$$

Now we construct a sequence $\Lambda^k \rightarrow \Lambda(z)$ such that Λ^k is feasible for the optimization problem (60) with z^k in place of z . Introduce vectors $C^k \in \mathbb{R}_+^I$ with $C_j^k = \sum_{i=I'+1}^I A_{ji} \Lambda_i(z^k)$. Put the first I' coordinates of Λ^k to be $\Lambda_i^k = (\Lambda_i(z) - \|C^k\|)^+ \wedge m_i z_i^k$, and the rest of them $\Lambda_i^k = \Lambda_i(z^k)$. That is, in the bandwidth allocation $\Lambda(z)$, the bandwidth C^k , which is required for the last $I - I'$ routes, is taken away from the first I' routes.

Since $z^k \rightarrow z$, $\Lambda^k \rightarrow \Lambda(z)$ and $\Lambda(z^k) \rightarrow \Lambda'$,

$$\sum_{i=1}^{I'} z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) \rightarrow l \quad \text{and} \quad \sum_{i=1}^{I'} z_i^k \mathcal{U}_i(\Lambda_i(z^k)/z_i^k) \rightarrow r.$$

Also, for all k ,

$$\sum_{i=I'+1}^I z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) = \sum_{i=I'+1}^I z_i^k \mathcal{U}_i(\Lambda_i(z^k)/z_i^k).$$

Then, by (61), for k big enough,

$$\sum_{i=1}^I z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) > \sum_{i=1}^I z_i^k \mathcal{U}_i(\Lambda_i(z^k)/z_i^k),$$

which contradicts to $\Lambda(z^k)$ being optimal for (60) with z^k in place of z . \square

PROOF OF COROLLARY 1. Fix an FMS $(\zeta, z)(\cdot)$. In §3, we discussed how Theorem 3 implies that $z(t) \rightarrow z^*$ as $t \rightarrow \infty$. Here we prove that $z(t) \rightarrow z^*$ implies $\zeta(t) \rightarrow \zeta^*$. It suffices to show that, for any $\varepsilon > 0$, there exists a t_ε such that, for all $t \geq t_\varepsilon$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i(t)(A) \leq \zeta_i^*(A^\varepsilon) + \varepsilon, \tag{62a}$$

$$\zeta_i^*(A) \leq \zeta_i(t)(A^\varepsilon) + \varepsilon, \tag{62b}$$

so fix $\varepsilon > 0$.

Downloaded from infoms.org by [131.155.2.68] on 17 September 2014, at 02:53 . For personal use only, all rights reserved.



For any $\delta \in (0, \min_{1 \leq i \leq I} z_i^*)$, there exists a τ_δ such that, for all $t \geq \tau_\delta$,

$$z^* - \delta := (z_1^* - \delta, \dots, z_I^* - \delta) \leq z(t) \leq (z_1^* + \delta, \dots, z_I^* + \delta) =: z^* + \delta.$$

Then, for all $t \geq s \geq \tau_\delta$ and i , we have

$$\underbrace{r_i(z^* - \delta, z^* + \delta)}_{=: r_i^\delta}(t-s) \leq \lambda_i(z^*)(t-s), \quad S_i(z, s, t) \leq \underbrace{R_i(z^* - \delta, z^* + \delta)}_{=: R_i^\delta}(t-s). \quad (63)$$

Recall from §3 that, for all i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i^*(A) = \eta_i \int_0^\infty \theta_i(A + (\lambda_i(z^*)s, s)) ds. \quad (64)$$

From the shifted fluid model Equation (5a) it follows that, for all $t \geq \tau_\delta$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i(t)(A) \leq \zeta_i(\tau_\delta)(\mathbb{R}_+ \times [t - \tau_\delta, \infty)) + \eta_i \int_{\tau_\delta}^t \theta_i(A + (S_i(z, s, t), t-s)) ds, \quad (65)$$

where, by (63), the second summand admits the estimate

$$\begin{aligned} \eta_i \int_{\tau_\delta}^t \theta_i(A + (S_i(z, s, t), t-s)) ds &\leq \eta_i \int_{\tau_\delta}^t \theta_i(A^{(R_i^\delta - r_i^\delta)(t-s)} + (\lambda_i(z^*)(t-s), (t-s))) ds \\ &= \eta_i \int_0^{t-\tau_\delta} \theta_i(A^{(R_i^\delta - r_i^\delta)s} + (\lambda_i(z^*)s, s)) ds. \end{aligned} \quad (66)$$

Take τ'_ε and $\delta \in (0, \min_{1 \leq i \leq I} z_i^*)$ such that

$$\eta_i \int_{\tau'_\varepsilon}^\infty \mathbb{P}\{D_i \geq s\} ds \leq \varepsilon/2, \quad (67)$$

$$\|R^\delta - r^\delta\| \tau'_\varepsilon \leq \varepsilon/2, \quad (68)$$

and take $t_\varepsilon \geq \tau_\delta + \tau'_\varepsilon$ such that, for all i ,

$$\zeta_i(\tau_\delta)(\mathbb{R}_+ \times [t_\varepsilon - \tau_\delta, \infty)) \leq \varepsilon/2. \quad (69)$$

Now, it follows from (65)–(66) by the choice of δ , τ_δ , τ'_ε and t_ε that, for all $t \geq t_\varepsilon$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\begin{aligned} \zeta_i(t)(A) &\stackrel{(69)}{\leq} \varepsilon/2 + \eta_i \int_0^{\tau'_\varepsilon} \theta_i(A^{(R_i^\delta - r_i^\delta)s} + (\lambda_i(z^*)s, s)) ds + \eta_i \int_{\tau'_\varepsilon}^\infty \theta_i(A^{(R_i^\delta - r_i^\delta)s} + (\lambda_i(z^*)s, s)) ds \\ &\stackrel{(68)}{\leq} \varepsilon/2 + \eta_i \int_0^{\tau'_\varepsilon} \theta_i(A^\varepsilon + (\lambda_i(z^*)s, s)) ds + \eta_i \int_{\tau'_\varepsilon}^\infty \mathbb{P}\{D_i \geq s\} ds \\ &\stackrel{(64), (67)}{\leq} \zeta_i^*(A^\varepsilon) + \varepsilon, \end{aligned}$$

i.e., we have (62a).

Similarly, we show (62b): for all $t \geq t_\varepsilon$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\begin{aligned} \zeta_i^*(t)(A) &\stackrel{(64)}{\leq} \eta_i \int_0^{\tau'_\varepsilon} \theta_i(A + (\lambda_i(z^*)s, s)) ds + \eta_i \int_{\tau'_\varepsilon}^\infty \mathbb{P}\{D_i \geq s\} ds \\ &\stackrel{(67)}{\leq} \eta_i \int_{t-\tau'_\varepsilon}^t \theta_i(A + (\lambda_i(z^*)(t-s), (t-s))) ds + \varepsilon/2 \\ &\stackrel{(63)}{\leq} \eta_i \int_{t-\tau'_\varepsilon}^t \theta_i(A^{(R_i^\delta - r_i^\delta)(t-s)} + (S_i(z, s, t), (t-s))) ds + \varepsilon/2 \\ &\stackrel{(68)}{\leq} \eta_i \int_{t-\tau'_\varepsilon}^t \theta_i(A^\varepsilon + (S_i(z, s, t), (t-s))) ds + \varepsilon/2 \\ &\stackrel{(5a)}{\leq} \zeta_i(t)(A^\varepsilon) + \varepsilon/2. \quad \square \end{aligned}$$

PROOF OF LEMMA 4. For all $s \leq t$ and $\varepsilon > 0$,

$$\int_s^t \mathbb{P}\{u+x \leq \xi < u+x' + \varepsilon\} du = \int_{s+x}^{t+x} \mathbb{P}\{\xi \geq u\} du - \int_{s+x'+\varepsilon}^{t+x'+\varepsilon} \mathbb{P}\{\xi \geq u\} du \leq \int_{s+x}^{s+x'+\varepsilon} \mathbb{P}\{\xi \geq u\} du \leq x' - x + \varepsilon.$$

The lemma follows as we first let $\varepsilon \rightarrow 0$ (applying the dominated convergence theorem) and then $s \rightarrow -\infty$, $t \rightarrow \infty$. \square

References

- [1] Asmussen S (2003) *Applied Probability and Queues*, 2nd ed. (Springer-Verlag, New York).
- [2] Ayesta U, Mandjes M (2009) Bandwidth-sharing networks under a diffusion scaling. *Ann. Oper. Res.* 170:41–58.
- [3] Balder EJ (2010) On subdifferential calculus. Lecture notes, http://www.staff.science.uu.nl/~balde101/cao10/cursus10_1.pdf.
- [4] Bonald T, Massoulié L (2001) Impact of fairness on Internet performance. *Proc. ACM Sigmetrics and Performance Conf., Boston*, (ACM, New York), 82–91.
- [5] Bonald T, Massoulié L, Proutière A, Virtamo J (2006) A queueing analysis of max–min fairness, proportional fairness and balanced fairness. *Queueing Systems* 53:65–84.
- [6] Borst S, Egorova R, Zwart B (2014) Fluid limits for bandwidth-sharing networks in overload. *Math. Oper. Res.* 39:533–560.
- [7] Bramson M (2005) Stability of networks for max–min fair routing. *Presentation at the 13th INFORMS Appl. Probab. Conf., Ottawa, ON, Canada*.
- [8] Chiang M, Shah D, Tang A (2006) Stochastic stability under network utility maximization: General file size distribution. *Proc. Annual Allerton Conf., Monticello, IL, USA*, 82–91.
- [9] de Veciana G, Lee T-L, Konstantopoulos T (1999) Stability and performance analysis of network supporting services with rate control—Could the Internet be unstable? *Proc. IEEE Infocom., New York*, 802–810.
- [10] de Veciana G, Lee T-L, Konstantopoulos T (2001) Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Trans. Networking* 9:2–14.
- [11] Egorova R, Borst SC, Zwart B (2007) Bandwidth-sharing networks in overload. *Performance Evaluation* 64:978–993.
- [12] Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence* (John Wiley & Sons, New York).
- [13] Gromoll HC, Williams RJ (2008) Fluid model for a data network with alpha-fair bandwidth sharing and general document size distributions: Two examples of stability. *IMS Collections—Markov Processes and Related Topics* 4:253–265.
- [14] Gromoll HC, Williams RJ (2009) Fluid limits for networks with bandwidth sharing and general document size distributions. *Ann. Appl. Probab.* 19:243–280.
- [15] Gromoll HC, Robert P, Zwart B (2008) Fluid limits for processor sharing queues with impatience. *Math. Oper. Res.* 33:375–402.
- [16] Jakubowski A (1988) Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probab. Math. Statist.* 9.1:95–114.
- [17] Kallenberg O (1983) *Random Measures* (Akademie-Verlag, Berlin).
- [18] Kang WN, Ramanan K (2012) Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Ann. Appl. Probab.* 22(2):477–521.
- [19] Kang WN, Kelly FP, Lee NH, Williams RJ (2009) State space collapse and diffusion approximation for a network operating under a fair bandwidth-sharing policy. *Ann. Appl. Probab.* 19:1719–1780.
- [20] Kelly FP, Williams RJ (2004) Fluid model for a network operating under a fair bandwidth-sharing policy. *Ann. Appl. Probab.* 14:1055–1083.
- [21] Kelly FP, Williams RJ (2010) Heavy traffic on a controlled motorway. Bingham NH, Goldie CM, eds. *Probab. Math. Genetics: Papers in Honour of Sir John Kingman* (Cambridge University Press, New York), 416–445.
- [22] Massoulié L (2007) Structural properties of proportional fairness: Stability and insensitivity. *Ann. Appl. Probab.* 17:809–839.
- [23] Massoulié L, Roberts JW (1999) Bandwidth sharing: Objectives and algorithms. *Proc. IEEE Infocom.* (Books in Statistics, New York), 1395–1403.
- [24] Mo J, Walrand J (2000) Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Networking* 8:556–567.
- [25] Reed J, Zwart B (2010) Limit theorems for bandwidth sharing networks with rate constraints. Preprint, revised March 19, 2013, New York University.
- [26] Roberts JW (2004) A survey on statistical bandwidth sharing. *Comput. Networks* 45:319–332.
- [27] Roberts JW, Massoulié L (1998) Bandwidth sharing and admission control for elastic traffic. *Proc. ITC Specialist Seminar, Yokohama, Japan*.
- [28] Takács L (1962) *Introduction to the Theory of Queues* (Oxford University Press, New York).
- [29] Ye H, Yao DD (2008) Heavy-traffic optimality of a stochastic network under utility-maximizing resource control. *Oper. Res.* 56:453–470.
- [30] Ye H, Yao DD (2010) Utility maximizing resource control: Diffusion limit and asymptotic optimality for a two-bottleneck model. *Oper. Res.* 58:613–623.