Closed queueing networks under congestion: non-bottleneck independence and bottleneck convergence

Jonatha Anselmi[1], Bernardo D'Auria[2], and Neil Walton[3]

[1]Basque Center for Applied Mathematics (BCAM), anselmi@bcamath.org
[2]Universidad Carlos III de Madrid, bernardo.dauria@uc3m.es
[3]University of Amsterdam, n.s.walton@uva.nl

Abstract

We analyze the behavior of closed multi-class product-form queueing networks when the number of customers grows to infinity and remains proportionate on each route (or class). First, we focus on the stationary behavior and prove the conjecture that the stationary distribution at non-bottleneck queues converges weakly to the stationary distribution of an ergodic, open product-form queueing network, which is geometric. This open network is obtained by replacing bottleneck queues with per-route Poissonian sources whose rates are uniquely determined by the solution of a strictly concave optimization problem. We strengthen such result by also proving convergence of the first moment of the queue lengths of non-bottleneck stations. Then, we focus on the transient behavior of the network and use fluid limits to prove that the amount of fluid, or customers, on each route eventually concentrates on the bottleneck queues only, and that the long-term proportions of fluid in each route and in each queue solve the dual of the concave optimization problem that determines the throughputs of the previous open network.

1 Introduction

Complex systems such as communication and computer networks are composed of a number of interacting particles (or customers) that exhibit important congestion phenomena as their level of interaction grows. The dynamics of such systems are affected by the randomness of their underlying events, e.g., arrivals of units of work, and can be described stochastically in terms of queueing network models. Provided that these are tractable, they allow one to make predictions on the performance achievable by the system, to optimize the network configuration and to perform capacity-planning studies. These objectives are usually difficult to achieve without a mathematical model because real systems are huge in size; e.g., Urgaonkar et al. [42].

The focus of this paper is on the well-known class of *closed* queueing network models introduced in Kelly [24] and Baskett et al. [4]. Specifically, a fixed number of customers circulate in a network following given *routes*. A route is a sequence of queues (or stations) that forms a cycle in the network. In terms of the amount of service required on each queue, users belonging to the same route are statistically equivalent. In contrast, users belonging to different routes can be statistically different. The stationary probability distribution of these queueing networks has the *product-form* property, which formally means that it can be written as the product of simple terms associated to each queue up to a normalizing constant. This surprising property represents a big step forward for the understanding of the stationary behavior of this queueing network model, as it drastically reduces the intrinsic computational complexity of solving the global balance equations of the underlying Markov chain. However, due to the quick growth of the state space and despite the attention devoted to this problem during the last decades, the computation of the normalizing constant remains a notoriously difficult task, especially when the number of customers is large. This issue limits the application of these models to optimization and dimensioning studies of real systems.

During the last decades, several approaches have been investigated to assess the stationary behavior of closed product-form queueing networks with a large number of customers. A large body of the literature aims at developing exact algorithms for the efficient computation of the normalizing constant or stationary performance indices such as mean queue lengths and throughputs; see, e.g., Reiser and Kobayashi [34], Reiser and Lavenberg [35], Harrison and Coury [19], Casale [9] and the references therein. To the best of our knowledge, no exact algorithm has a running time that is polynomial with respect to the numbers of routes, customers and queues. Motivated by this difficulty, a number of alternative analyses emerged in the literature for the stationary behavior. These mainly consist in:

- Using the mean value analysis (MVA) by Reiser and Lavenberg [35] to develop iterative or fixed-point algorithms; e.g., Schweitzer [37], Chandy and Neuse [11], Pattipati et al. [31], Wang et al. [44]. While these techniques improve the running time of MVA, there is no guarantee that they converge to the exact solution except for particular cases.

- Developing efficient bottleneck identification techniques; e.g., Schweitzer [38], Schweitzer et al. [39], Casale and Serazzi [10], Anselmi and Cremonesi [1]. This approach aims at reducing the network size by ignoring the impact of the stations that have a minor influence on the overall performance.
- Studying the network behavior when some parameter approaches a limiting value of practical interest, which is the approach followed in this article.

A number of limiting regimes have been studied in queueing networks. For the case of single-class networks, see Goodman and Massey [17]. More generally, through a number of examples, Whitt [45] expounds the technique of approximating closed systems with the behaviour of an appropriate open queueing system. For open and closed migration processes, Pollett [33] and Brown and Pollett [8] find methods to identify bottlenecks and, in addition, characterize the total variation distance between the output process of some queue and a Poisson process. Works analyse the fluid and heavy-traffic limit of single-class closed queueing networks. Chen and Mandelbaum [13] provides a heavy traffic approximation of closed networks, and a characterization of bottleneck behaviour is found. The paper extends these results to two classes of customers with different priorities. In Kaspi and Mandelbaum [20], the generality of the service type distributions is greatly increased and a concrete connection between the limiting behaviour of bottleneck queues and the stationary distributions of a Brownian network is made. The book of Chen and Yao [12] analyzes the fluid and diffusion limits for generalized Jackson networks, for both open and closed topologies. The word generalized refers to the fact that the external Poisson arrivals are substituted by general and independent renewal processes. The analysis there, though, mainly considers single-class networks, or in the multi-class case focuses on the FIFO discipline. This discipline is very different from the processor sharing one and shows qualitatively different behavior, for instance unexpected instability occurs when all stations of the network are individually stable for the solution of the traffic flow equations; Kumar and Seidman [27]. Fluid analysis of multi-class open queueing networks is found in Bramson [7]. Detailed discussions on Bramson's work will be made in subsequent sections of this article.

In the closed multi-class queueing networks investigated in this article, a number of works have assumed the existence of an $M/M/\infty$ queue and allowed the total number of jobs, say n, grow in proportion to its service (also known as 'think') times; see e.g. McKenna and Mitra [30], McKenna [29], Berger et al. [5]. Further analyses let n grow to infinity in proportion to the number of stations; Knessl and Tier [25, 26]. Finally, another important approach is to study the network behavior when n grows to infinity keeping fixed the proportions of customers in each route; Pittel [32], Schweitzer [38], Balbo and Serazzi [3], Walton [43], Walton et al. [22], Anselmi and Cremonesi [1], George et al. [16].

The focus of this paper is on the last approach described above where n, the population vector, grows to infinity keeping fixed the proportions of customers in each route. In this limiting regime, it is known that some queues, called *bottlenecks* in the following, increase their backlog proportionally to n, see Pittel [32] and Walton et al. [22], and uniquely determine the throughput of customers along each route by means of a concave optimization problem; Schweitzer [38], Walton [43]. Interestingly, this optimization problem coincides with the utility optimization problem that determines the fractions of bandwidth (or rates) allocated to multiple classes of concurrent Internet flows (or end-to-end Internet connections); see Kelly et al. [23], Srikant [41]. On the other hand, the amount of backlog in each *non*-bottleneck is strictly bounded by $O(n)$ but in general its limiting behavior is not known. There is numerical evidence to support the conjecture that the limiting stationary distribution of each non-bottleneck queue is geometric; Balbo and Serazzi [3]. Such convergence was proved by Pittel [32] under the assumption that the mode of the stationary distribution is unique; see also Gordon and Newell [18], Lipsky et al. [28], Anselmi and Cremonesi [1]. This assumption, for instance, is not satisfied when the number of bottlenecks is less than the number of routes but greater than one, which is a natural scenario in real networks. As we shall explain, one of the principal contributions of this paper is to prove this conjecture by removing the unimodal assumption.

The above conjecture provides an accurate and efficient approximation to mean performance indices that takes into account the contribution of non-bottlenecks. The impact on performance of non-bottlenecks becomes non-negligible if their number prevails significantly, and this is often the case; see Casale and Serazzi [10], Anselmi and Cremonesi [1]. This approximation allows for the direct development of efficient optimization frameworks able to address, for instance, data-center consolidation problems [2], where the objective is to reduce the cost and the size of a data-center while guaranteeing a given performance level. Furthermore, it provides a good initial guess for the iterative or fixed-point algorithms mentioned above.

1.1 Our contribution.

Following the approach considered by a large body of the literature, we are interested in the behavior of closed product-form queueing networks when the number of customers in each route grows to infinity proportionally. This is mainly motivated because real networks are populated by a large number of customers; e.g., Urgaonkar et al. [42]. Our objective is two-fold and consists in analyzing the problem from two contrasting standpoints.

The first part of this paper focuses on the *stationary* behavior of these networks. We prove the conjecture that the stationary distribution of non-bottlenecks converges weakly to the stationary distribution of an ergodic, open product-form queueing network, with geometrically distributed queue lengths. This open network is obtained by replacing bottlenecks with per-route Poissonian sources whose rates are uniquely determined by the solution of a strictly concave optimization problem. In addition, we strengthen such result by proving that the mean per-route number of customers of non-bottlenecks converges to the corresponding mean number of customers of the same open network, which is important from an operational standpoint.

On the other hand, the second part of this paper focuses on the *transient* behavior of the fluid limit version of the network. We start from any arbitrary distribution of customers in the network that preserves the proportional distribution. Then, we let their total number go to infinity. This allows, after a renormalization and a limit procedure, to construct the fluid limit version of the network with a given initial concentration of fluid in its queues. We prove that the amount of fluid, or customers, on each route eventually concentrates, as time increases, on the bottleneck queues only and that the (long-term) proportions of fluid in each route and in each bottleneck solve the dual of the concave optimization problem that determines the throughputs of the open network described in the first part above. Our proof for closed queueing networks uses an entropy Lyapunov function similar to the one used by Bramson [7] to establish convergence properties of the fluid-limit equations of open queueing networks.

The technical difference behind the two results above is the order in which the limit in the number of customers and the limit in time are taken. In the second part, the limit in the number of customers is taken *before* the limit in time, and vice versa. In stochastic systems, these two limits do not commute in general, but for the class of queueing networks investigated in this paper we prove that they do. Taking the limit in the number of customers first provides a natural way to look at the evolution of a network populated by a large number of customers and, by subsequently taking a limit in time, we justify fluid model arguments within the queueing literature. The second result proven in this paper, thus, increases the robustness of the approach taken in the first part, which has been followed by several researchers as referenced above. Furthermore, it can be also seen as a queueing theoretic analysis of the utility optimization found in congestion control protocols; e.g., Kelly et al. [23] and Srikant [41].

1.2 Organization.

In Section 2, we introduce the model considered in this paper. In particular, we include two Markov descriptions of a closed queueing network, relevant quantities such as bottleneck and non-bottleneck queues are defined, and an expression for a fluid model of a closed queueing network is given. In Section 3, we present the three main results of this paper: Section 3.1 shows the asymptotic independence of non-bottleneck queues in the large-population limit; Section 3.2 shows the convergence of the Markov closed queueing network to a fluid solution; and Section 3.3 states that a fluid solution converges to the set of bottlenecks in a way that minimizes a certain entropy Lyapunov function. In Section 4, we prove the main results stated in Section 3. We respectively prove the results of Sections 3.1, 3.2 and 3.3 in Sections 4.1, 4.2 and 4.3.

2 Closed queueing networks models.

We consider closed, multi-class queueing networks in the sense of Kelly [24] and Baskett et al. [4]. The set $\mathcal{J} \subset \mathbb{N}$ denotes the set of queues (or stations) and we let $J = |\mathcal{J}|$. The set $\mathcal{I} \subset 2^{\mathcal{J}}$ denotes the set of routes (or classes) and we let $I = |\mathcal{I}|$. A route is a sequence of queues visited by a customer during one cycle of the network. We assume, for simplicity, that each customer visits each queue at most once within a cycle of the network. Within each route $i = \{j_1^i, ..., j_{k_i}^i\}$, we associate a route order $(j_1^i, ..., j_{k_i}^i)$. For $k = 1, ..., k_i - 1$, a customer departing queue j_k^i will next join queue j_{k+1}^i and a customer departing queue $j_{k_i}^i$ will join queue queue j_1^i. Unless otherwise specified, i will be used to index routes and j will be used to index queues. We assume that a constant number of customers circulate along each route of the network. We denote by $n = (n_i : i \in \mathcal{I}) \in \mathbb{N}^I$ the *population vector*, i.e., the total number of customers on each route. When joining queue j, we assume that route-i customers require amounts of service that are independent and exponentially distributed with mean μ_{ji}^{-1}. At each queue, we assume that customers are served at rate 1 according to a processor sharing discipline and customers joining a queue take a position uniformly at random in the queue. Thus, if a route-i customer does not join a queue j, i.e. $j \notin i$, we may assume $\mu_{ji}^{-1} = 0$.

2.1 Two Markov models of closed queueing networks.

Firstly, we could describe the exact location of each customer in the queue according to its route type. Here the *explicit state* of a queue j would be a vector $s_j = (i_j(k) : k = 1, ..., m_j)$, where m_j is the number of customers in the queue and $i_j(k)$ is the route type of the customer in the k^{th} position. The explicit state of the network would then be the vector of each queue's state, $s = (s_j : j \in \mathcal{J})$. We then let $\mathcal{X}(n)$ be the set of the explicit states where the number of customers of each route type i is n_i.

Secondly, we could ignore positional information about customers within a queue, and instead, just consider the number of each route type at a queue. Here, we let $m = (m_{ji} : j \in \mathcal{J}, i \in \mathcal{I}, j \in i)$ be a network *state*, where m_{ji} represents the number of route i customers in queue j. Thus, $\mathcal{S}(n) = \{m : \sum_j m_{ji} = n_i, i \in \mathcal{I}\}$ is the state space of this Markov chain. It may be verified in a straightforward manner that this state space is finite and irreducible.

Under the above assumptions, it is known (Kelly [24] and Baskett et al. [4]) that the stationary distribution of being in state s, which we denote by $\pi(s|n)$, or in state m, which we denote $\pi(m|n)$, is respectively

$$\pi(s|n) = \frac{1}{B(n)} \prod_{j\in\mathcal{J}} \prod_{k=1}^{m_j} \mu_{j i_j(k)}^{-1} = \frac{1}{B(n)} \prod_{j\in\mathcal{J}} \prod_{i:j\in i} \mu_{ji}^{-m_{ji}}, \qquad s \in \mathcal{X}(n), \tag{2.1}$$

$$\pi(m|n) = \frac{1}{B(n)} \prod_{j\in\mathcal{J}} \left(\binom{m_j}{m_{ji} : i \ni j} \prod_{i:j\in i} \mu_{ji}^{-m_{ji}} \right), \qquad m \in \mathcal{S}(n), \tag{2.2}$$

where $B(n)$ is the normalizing constant

$$B(n) \stackrel{\text{def}}{=} \sum_{m\in\mathcal{S}(n)} \prod_{j\in\mathcal{J}} \left(\binom{m_j}{m_{ji} : i \ni j} \prod_{i:j\in i} \mu_{ji}^{-m_{ji}} \right), \tag{2.3}$$

that is the same in (2.1) and (2.2); see [4]. To count the possible orderings of customers inside a queue, we use the multinomial coefficient

$$\binom{m_j}{m_{ji} : i \ni j} \stackrel{\text{def}}{=} \frac{m_j!}{\prod_{i:j\in i} m_{ji}!}.$$

In the following, we mainly consider the stationary distribution (2.2), while the expression (2.1) will be used in proofs.

Remark 1. *There are a large number of generalizations of our queueing network where* (2.2) *still gives the stationary probability of the per-route number of customers in each queue. For instance, we could generalize our processor-sharing discipline to the class of symmetric queueing disciplines with unit service capacity, see Kelly [24, Section 3.3]. We could also generalize service requirements to be independent with mean* μ_{ji}^{-1} *and having a rational Laplace transform. If we keep the assumption that service requirements are exponential and we assume that the state space of our Markov description is irreducible, then we could generalize to allow any service discipline that allocates service amongst customers in a way that does not discriminate between the route types of customers at the queue. Such service disciplines are described by Kelly [24, Section 3.1]. Finally, in all cases, we may elaborate the routing structure of our network; we may allow a route* i *customer to be routed through the network as a Markov chain (with finite expected exit time) whose states are determined by the set of queues so far visited.*

Although generalizations leading to stationary distribution (2.2) are abundant (as explained in the above remark), we do not explore these in further detail for simplicity of exposition. Thus, our results in Theorems 1 and 2 generalize to such cases. On the other hand, Theorems 3 and 4 apply to the specific case of processor-sharing queues, and thus they do not apply directly to all the generalizations described above. The extension of Theorems 3 and 4 when other service disciplines are considered, e.g., last-come-first-served preemptive-resume, is left as future work.

2.2 Throughputs and bottlenecks.

A key quantity of interest is the rate at which customers complete service on each route with respect to a reference queue (say j_1^i for route i). Let

$$\Lambda_i(n) \stackrel{\text{def}}{=} \sum_{\substack{m\in\mathcal{S}(n):\\ m_{j_1^i}>0}} \mu_{j_1^i i} \frac{m_{j_1^i i}}{m_{j_1^i}} \pi(m|n) \tag{2.4}$$

be the *throughput* of customers on route i observed at queue j_1^i. In addition, let

$$U_j(n) \stackrel{\text{def}}{=} \sum_{i:j\in i} \frac{\Lambda_i(n)}{\mu_{ji}} \tag{2.5}$$

be the *utilization* (or load) of queue j. Using Little's law, $U_j(n)$ can be interpreted as the proportion of time where queue j is busy.

Remark 2. *The expression* (2.4) *gives the service completion rate of route* i *customers at queue* j_1^i*, i.e.,* $\mu_{j_1^i i}$*, times the service rate devoted to these customers at that queue, i.e.,* $m_{j_1^i i}/m_{j_1^i}$*, times the stationary probability that there are* m *customers in the network, i.e.,* $\pi(m|n)$*. Thus, this expression gives the mean rate (or throughput) for which route* i *customers leave queue* j_1^i*. Note the route* i *throughputs at each queue* $j \in i$ *must be equal. Without loss of generality, we considered queue* j_1^i*.*

A more concise expression for the per-route throughput is given in the following lemma; see, e.g., Bolch et al. [6, Formula (8.28)] or, for single class networks, Chen and Yao [12, Formula (2.10)].

Lemma 1.

$$\Lambda_i(n) = \frac{B(n-e_i)}{B(n)}, \qquad i \in \mathcal{I} \tag{2.6}$$

where $B(n)$ *is the normalizing constant* (2.3) *and* e_i *is the* i^{th} *unit vector in* $\mathbb{R}_+^I$.

In agreement with other works (e.g., Anselmi and Cremonesi [1], Balbo and Serazzi [3]), we define a bottleneck queue as a queue whose service is saturated.

Definition 1. *Queue $j \in \mathcal{J}$ is called bottleneck if and only if*

$$\lim_{c\to\infty} U_j(cn + u_c) = 1, \tag{2.7}$$

where $\{u_c\}_{c\in\mathbb{N}}$ is any uniformly bounded sequence such that $cn + u_c \in \mathbb{Z}_+^I$. We define the set $\bar{\mathcal{J}} \subseteq \mathcal{J}$ to be the set of bottlenecks and let $\bar{J} = |\bar{\mathcal{J}}|$. Similarly, we define $\mathcal{J}^\circ = \mathcal{J}\backslash\bar{\mathcal{J}}$ to be the set of non-bottlenecks and let $J^\circ = |\mathcal{J}^\circ|$.

We will think of $\mathcal{J}^\circ$ as the 'open part' of the closed network under investigation.

2.3 Non-bottleneck queues and open queueing networks.

We are interested in the probability distribution of non-bottleneck queues. For this reason, we consider queue-size vectors $m^\circ = (m^\circ_{ji} : j \in \mathcal{J}^\circ, i \in \mathcal{I}, j \in i) \in \mathbb{Z}_+^{J^\circ \times I}$ and

$$\pi^\circ(m^\circ|n) \stackrel{\text{def}}{=} \sum_{\substack{\bar{m}\in\mathbb{Z}_+^{\bar{\mathcal{J}}}:\\ (\bar{m},m^\circ)\in\mathcal{S}(n)}} \pi((\bar{m}, m^\circ)|n), \qquad j \in \mathcal{J}^\circ, i \in \mathcal{I}, \tag{2.8}$$

which defines the stationary probability that non-bottleneck queues are in state m°.

We also define

$$\pi^\circ_\Lambda(m^\circ) \stackrel{\text{def}}{=} \prod_{j\in\mathcal{J}^\circ} \pi^\circ_{j,\Lambda}(m^\circ), \tag{2.9a}$$

$m^\circ \in \mathbb{Z}_+^{J^\circ\times I}$, $\Lambda \in \mathbb{R}_+^I$, where

$$\pi^\circ_{j,\Lambda}(m^\circ) \stackrel{\text{def}}{=} \left(1 - \sum_{i:j\in i} \frac{\Lambda_i}{\mu_{ji}}\right) \binom{m_j}{m_{ji} \;:\; i \ni j} \prod_{i:j\in i} \left(\frac{\Lambda_i}{\mu_{ji}}\right)^{m_{ji}}. \tag{2.9b}$$

The distribution π, given by (2.2), refers to the stationary distribution of a closed queueing network. In contrast, the distribution π°_Λ can be shown to be the stationary distribution of an *open* queueing network constructed on queues $\mathcal{J}^\circ$. Here customers arrive on each route as a Poisson process of rate $\Lambda = (\Lambda_i : i \in \mathcal{I})$ and depart the network after receiving service at each queue on their route $i \cap \mathcal{J}^\circ$, see Kelly [24] and Baskett et al. [4].

2.4 Fluid Model.

In order to study the transient behavior of our closed queueing network, we will analyze the following fluid model.

Definition 2 (Closed queueing network fluid model)**.** *The processes $m(t) = (m_{ji}(t) : j \in \mathcal{J}, i \in \mathcal{I}, j \in i)$ and $L(t) = (L_{ji}(t) : j \in \mathcal{J}, i \in \mathcal{I}, j \in i)$ form a* fluid solution *(or fluid limit) of our closed queueing network if they satisfy the following conditions:*

$$m_{j^i_k i}(t) = L_{j^i_{k-1} i}(t) - L_{j^i_k i}(t), \tag{2.10a}$$

$$\sum_{i:j\in i} \frac{1}{\mu_{ji}} (L_{ji}(t) - L_{ji}(s)) \le t - s, \tag{2.10b}$$

$$L_{ji}(t) \text{ is increasing,} \tag{2.10c}$$

$$\text{if } m_j(t) > 0 \text{ then } \Lambda_{ji}(t) = \frac{m_{ji}}{m_j}\mu_{ji}, \tag{2.10d}$$

$$\sum_{j:j\in i} m_{ji}(t) = n_i. \tag{2.10e}$$

where we used the definition $\Lambda_{ji}(t) \stackrel{\text{def}}{=} dL_{ji}(t)/dt$. Here, $j \in \mathcal{J}, i \in \mathcal{I}, j \in i$, $t \ge s \ge 0$. Also, j^i_k is the k^{th} queue on route i and j^i_{k-1} is the queue before the k^{th} queue (we use the convention that the queue before j^i_1 is $j^i_{k_i}$).

The (j, i)-component of the process $m(t)$ denotes the amount of fluid of type i contained at time t at queue $j \in i$. The process $\Lambda_{ji}(t)$ gives the instantaneous throughput of this type of fluid at the same queue, while the process $L_{ji}(t)$ gives the total amount of fluid of this type flowed out from queue j by time t.

The conditions (2.10), thus, are the defining properties of a fluid solution and we observe that they are analogous to the ones used by Bramson [7, Formulas (2.3)-(2.6)], which hold for the open versions of the considered closed queueing networks[1]. In particular, conditions (2.10a), (2.10c), (2.10e) are basic and

[1] Actually, the open queueing networks defined in Bramson [7], called head-of-the-line processor sharing networks, appear different from ours. In fact, upon completion of service at one queue, a class-i_1 customer becomes of class i_2 with probability $p_{i_1 i_2}$. However, one can easily build a mapping from one representation to the other.

relate queue lengths in an obvious manner, condition (2.10d) is the property that defines a processor-sharing discipline, and condition (2.10b) takes into account the maximal processing rate of the system.

We note that the condition (2.10b) implies that L_{ji} is Lipschitz continuous. By (2.10a), this is also true for m_{ji}. Lipschitz continuity implies absolute continuity, and therefore the processes L_{ji} and m_{ji} are differentiable almost everywhere with respect to the Lebesgue measure. Throughout this document, the term *for almost every* will refer to a set of real numbers whose complement has Lebesgue measure zero. Shortly, we will prove that the limit of the closed queueing network described previously satisfies the fluid model (2.10).

3 Results on closed queueing networks.

In this section, we present the main results of this article. Namely, Theorems 1 and 2, which state the independence of the per-route numbers of customers in non-bottlenecks and the convergence of their expectations under a large-population limit (Section 3.1); Theorem 3, which states that the stochastic process limit of a closed queueing network is a solution to the fluid equations (2.10) (Section 3.2); Theorem 4, which states the convergence in time of the fluid model (2.10) via a Lyapunov function argument (Section 3.3).

3.1 Independence of non-bottleneck queues.

We will demonstrate that non-bottleneck queues become independent as our closed queueing network becomes congested. We now consider the limiting behavior of our queueing network with cn customers when $c \to \infty$. Given the constraints on the number of customers on each route, the random variables of the stationary number of per-route customers in each queue are certainly dependent; however, in classic open queueing networks, where all customers arrive from an external source and eventually leaves the network, they are shown to be independent, see Kelly [24] and Baskett et al. [4].

We consider our stationary distributions for closed multiclass queueing networks, (2.8), and open multiclass queueing networks, (2.9). In informal terms, we wish to prove that for each $n \in \mathbb{R}_+^I, m \in \mathbb{Z}_+^{J^\circ \times I}$

$$\pi^\circ(m^\circ|cn) \xrightarrow[c\to\infty]{} \pi^\circ_{\Lambda^*}(m^\circ). \tag{3.1}$$

To prove such a statement, we must identify the throughput vector $\Lambda^* \in \mathbb{R}_+^I$ and the set of non-bottleneck queues $\mathcal{J}^\circ$. Since the vector cn need not to belong to $\mathbb{Z}_+^I$, we consider $cn+u_c$ where $\{u_c\}_{c\in\mathbb{N}}$ is any uniformly bounded sequence such that $cn + u_c \in \mathbb{Z}_+^I$. Our theorem can then be written as follows.

Theorem 1. *For* $n \in \mathbb{R}_+^I, m^\circ \in \mathbb{Z}_+^{J^\circ \times I}$

$$\pi^\circ(m^\circ|cn + u_c) \xrightarrow[c\to\infty]{} \pi^\circ_{\Lambda^*(n)}(m^\circ) \tag{3.2}$$

where $\Lambda^*(n) = (\Lambda_1^*(n), \dots, \Lambda_I^*(n))$ *is the unique optimizer of the strictly-concave optimization problem*

$$\textit{maximize} \quad \sum_{i\in\mathcal{I}} n_i \log \Lambda_i \tag{3.3a}$$

$$\textit{subject to} \quad \sum_{i:j\in i} \Lambda_i/\mu_{ji} \le 1, \qquad j \in \mathcal{J} \tag{3.3b}$$

$$\textit{over} \quad \Lambda_i \ge 0, \qquad i \in \mathcal{I}, \tag{3.3c}$$

and where $\mathcal{J}^\circ$*, the set of non-bottlenecks, is given by the set of queues* j *such that*

$$\sum_{i:j\in i} \frac{\Lambda_i^*(n)}{\mu_{ji}} < 1.$$

The solution of optimization problem (3.3), $\Lambda^*(n)$, is known in the literature as the *proportionally-fair* allocation (see Kelly [21]) and, interestingly, emerged independently as a model for the sharing of bandwidth among Internet connections (e.g., Srikant [41]). For a detailed treatment of the relationship between closed queueing networks and the proportionally-fair allocation, see Walton [43].

Theorem 1 proves that non-bottleneck queue lengths converge in distribution. Unfortunately, this is not enough to claim that the first moments converge as well. Convergence of the first moments is important from a practical standpoint, particularly for mean value analyses. This result is demonstrated in the next theorem. For our closed queueing network with n customers, let $M^\circ(n) \in \mathbb{Z}_+^{J^\circ \times I}$ be the stationary number of customers from each route at each non-bottleneck queue. In other words, $\Pr(M^\circ(n) = m^\circ) = \pi^\circ(m^\circ|n)$. Similarly, let M°_Λ be the stationary number of customers from each route at each queue in the open queueing network (2.9a), i.e., $\Pr(M^\circ_\Lambda = m^\circ) = \pi^\circ_\Lambda(m^\circ)$.

Theorem 2.

$$EM^\circ(cn + u_c) \xrightarrow[c\to\infty]{} EM^\circ_{\Lambda^*(n)} \tag{3.4}$$

where $\Lambda^*(n) = (\Lambda_1^*(n), \dots, \Lambda_I^*(n))$ *is the unique optimizer of* (3.3).

3.2 Existence of fluid limits for closed queueing networks.

In Section 3.1, we analyzed the *stationary* probability distribution (2.2) of the closed queueing networks under investigation in the large-population limit. Now, we focus on the *transient* probability distribution in the large-population limit and then study the evolution in time of the system. In other words, the limit in time is now taken after the limit in the number of customers. In stochastic systems it is known that both limits are not interchangeable in general. The fluid limit, see Definition 2, is a natural framework that allows for the analysis of such scenario.

We consider a sequence of the closed queueing networks as described in Section 2. In this sequence, the only variables that change are the number of customers on each route (the number of queues, routes, service distributions are kept fixed). We let the vector $n \in \mathbb{R}_+^I$ be the proportion of customers on each route of the network. In the c^{th} network of this sequence of closed queueing networks, there are $cn + u_c$ customers on each route, where $\{u_c\}_{c\in\mathbb{N}}$ is a bounded sequence of variables in $\mathbb{R}_+^I$ such that $cn + u_c \in \mathbb{Z}_+^I$.

We let $M_{ji}^c(t)$ be the number of route-i customers in queue j at time t of the c^{th} closed queueing network. We let $L_{ji}^c(t)$ give the total number of route-i customers served by queue j by time t in the c^{th} closed queueing network. From this, we define the rescaled processes

$$\bar{M}_{ji}^c(t) \stackrel{\text{def}}{=} \frac{M_{ji}^c(ct)}{c}, \qquad \bar{L}_{ji}^c(t) \stackrel{\text{def}}{=} \frac{L_{ji}^c(ct)}{c}, \qquad j \in \mathcal{J},\ i \in \mathcal{I},\ j \in i.$$

We wish to show that the vector processes $\bar{M}^c$ and $\bar{L}^c$ converge to a fluid solution.

Theorem 3. *The sequence of stochastic processes $\{(\bar{M}^c, \bar{L}^c\}_{c\in\mathbb{N}}$ converges weakly to a continuous process (m, L) that satisfies the fluid solution equations* (2.10).

3.3 Convergence of the fluid solution.

Having established the existence of a fluid solution of our closed queueing networks, (m, L), now our goal is to study its evolution in the long-term, i.e., $m(t)$ when $t \to \infty$. In our main result, we show that the amount of fluid m eventually concentrates on the bottleneck queues only and that the long-term proportions of fluid in each route and in each queue solve the dual of optimization problem (3.3).

From the stationary distribution $\pi(m|n)$, given by (2.2), and under the same premise of Theorem 1, we can show that

$$\lim_{c\to\infty} \frac{1}{c} \log \pi(cm|cn) = -\beta(m)$$

where

$$\beta(m) \stackrel{\text{def}}{=} \sum_{j\in\mathcal{J}} \sum_{\substack{i:j\in i\\ m_{ji}>0}} m_{ji} \log \frac{m_{ji}\mu_{ji}}{m_j} \tag{3.5}$$

$$= \sum_{\substack{j\in\mathcal{J}:\\ m_j>0}} m_j \sum_{i:j\in i} p_{ji} \log \frac{p_{ji}}{\mu_{ij}^{-1}}. \tag{3.6}$$

Here, we have used the notation $p_{ji} \stackrel{\text{def}}{=} m_{ji}/m_j$.

Remark 3 (Relative Entropy)**.** *A key quantity that will be useful in our proofs is the unnormalized relative entropy*[2]

$$D(p||q) \stackrel{\text{def}}{=} \sum_i p_i \log \frac{p_i}{q_i}, \qquad p, q \in \mathbb{R}_+^I.$$

We notice that the function β, given by (3.6)*, is a linear combination of these unnormalized relative entropies for each queue. Furthermore, we note that if $\sum_i p_i = \sum_i q_i$, then we can renormalize and treat p and q as probability distributions. In this case, one can show that $D(p||q) \geq 0$, using Jensen's inequality, and that $D(p||q)$ is minimized if $p = q$ where its value is* 0.

Remark 4. *It is an interesting observation that the rate of decrease of $\beta(m)$, the (negative) entropy of states, will be determined by the relative entropy between the rates of service; see Proposition 7 below. Although distinct, this argument is similar to those argued for Markov processes by Spitzer [40] and more recently by Dupuis and Fischer [15]. A further important reference is Bramson [7], which shows, for the open versions of the closed queueing networks considered in this paper, that the amount of fluid in non-bottlenecks converges to zero.*

In Bramson [7], the main argument behind the fluid convergence to zero for non-bottleneck stations relies on closing *the open network with an additional (artificial) queue whose service rate equals the overall throughput of the network. All customer routes enter this additional queue and it is assumed that this queue never empties. This plays a significant role in regulating traffic to match that of the open network. So although a closed queueing network appears in that analysis, this additional queue introduces an important loss of generality.*

[2] This entropy is unnormalized because we do not enforce the condition that its arguments are probability distributions.

The function $\beta(m)$ forms a natural candidate for a Lyapunov function. We need to show that, for $n \in \mathbb{R}_+^I$ fixed, $\beta(m(t))$ decreases to its minimal value

$$\beta^* \stackrel{\text{def}}{=} \text{minimize} \quad \beta(m) \quad \text{subject to} \quad \sum_{j \in i} m_{ji} = n_i, \quad i \in \mathcal{I}. \tag{3.7}$$

We show in Lemma 11 that the above optimization problem is the dual of the problem (3.3), and considering the set of points attaining this minimal value

$$\mathcal{M}(n) \stackrel{\text{def}}{=} \text{argmin} \quad \beta(m) \quad \text{subject to} \quad \sum_{j \in i} m_{ji} = n_i, \quad i \in \mathcal{I}, \tag{3.8}$$

we have that when the network is in one of these states the corresponding throughputs are uniquely given by the $\Lambda^*(n)$ vector given in Theorem 1.

The following theorem shows how fluid eventually distributes among network stations.

Theorem 4. *Let $(m(t), L(t))$ be a solution of the fluid model* (2.10) *with $m(0)$ satisfying equation* (2.10e) *for a given vector n, then,*

$$\beta(m(t)) \searrow \beta^*, \qquad \text{as} \qquad t \to \infty$$

and, moreover,

$$\min_{m^* \in \mathcal{M}(n)} |m(t) - m^*| \to 0, \qquad \text{as} \qquad t \to \infty$$

where the fluid point $m^* \in \mathcal{M}(n)$ induces, via equation (2.10d)*, the same throughputs $\Lambda^*(n)$ given in Theorem 1.*

In the above theorem, the norm $|m|$ is the Euclidean norm in $\mathbb{R}^{\mathcal{I} \times \mathcal{J}}$.

4 Proofs of main results.

We now focus on proving the main results of this paper, namely, Theorems 1 and 2 (Section 4.1), Theorem 3 (Section 4.2) and Theorem 4 (Section 4.3).

4.1 Analysis of non-bottleneck queues.

First, we develop a proof of Theorem 1. Recalling the informal statement (3.1), we need to verify that for each $n \in \mathbb{R}_+^I, m \in \mathbb{Z}_+^{J^\circ \times I}$

$$\pi^\circ(m^\circ|cn) \xrightarrow[c \to \infty]{} \pi^\circ_{\Lambda^*}(m^\circ), \tag{4.1}$$

for some $\Lambda^* \in \mathbb{R}_+^I$ for some set of non-bottleneck queues $\mathcal{J}^\circ$. Before verifying such statement, we must identify the relevant throughput vector Λ and non-bottleneck queues J°. The following result, which was proven by Walton [43], characterizes Λ^* and $\mathcal{J}^\circ$.

Proposition 1 ([43])**.**

$$\Lambda_i^*(n) = \lim_{c \to \infty} \Lambda_i(cn + u_c), \qquad i \in \mathcal{I}$$

where $\Lambda_i^*(n)$ is the unique minimizer of the concave optimization problem (3.3) *and so that $\Lambda_i(cn + u_c)$ is defined on a point in its domain, $\{u_c\}_{c \in \mathbb{N}}$ is any bounded sequence in $\mathbb{R}_+^I$ such that $cn + u_c \in \mathbb{Z}_+^I$.*

Consequently, $j \in \mathcal{J}^\circ$ if and only if

$$\sum_{i:j \in i} \frac{\Lambda_i^*(n)}{\mu_{ji}} < 1.$$

Now, we show that (3.2) holds. Before proceeding with this proof, we introduce a specific closed queueing network, which will helps us in the proof. Recall the closed queueing network defined on set $\mathcal{J}$ introduced in Section 2. Consider a queueing network defined exactly as in Section 2, except that queues $\mathcal{J}^\circ$ are removed. The resulting queueing network has states $\bar{m} = (\bar{m}_{ji} : j \in \bar{\mathcal{J}}, i \in \mathcal{I}, j \in i) \in \mathbb{Z}_+^{\bar{J} \times I}$; if there are $n \in \mathbb{Z}_+^I$ customers on each route, this network has state space $\bar{\mathcal{S}}(n) = \{\bar{m} : \sum_j \bar{m}_{ji} = n_i, i \in \mathcal{I}\}$, stationary distribution

$$\bar{\pi}(\bar{m}|n) = \frac{1}{\bar{B}(n)} \prod_{j \in \bar{\mathcal{J}}} \left(\binom{\bar{m}_j}{\bar{m}_{ji} : j \in i} \prod_{i:j \in i} \mu_{ji}^{-\bar{m}_{ji}} \right), \qquad \bar{m} \in \bar{\mathcal{S}}(n), \tag{4.2}$$

where

$$\bar{B}(n) = \sum_{\bar{m} \in \bar{\mathcal{S}}(n)} \prod_{j \in \bar{\mathcal{J}}} \left(\binom{\bar{m}_j}{\bar{m}_{ji} : j \in i} \prod_{i:j \in i} \mu_{ji}^{-\bar{m}_{ji}} \right),$$

and stationary throughput

$$\bar{\Lambda}(n) = \frac{\bar{B}(n - e_i)}{\bar{B}(n)}.$$

Proposition 1 holds for this network and, here, we would consider $\bar{\Lambda}^*(n)$, the solution to the optimization

$$\text{maximize} \quad \sum_{i\in\mathcal{I}} n_i \log \Lambda_i \quad \text{subject to} \quad \sum_{i:j\in i} \Lambda_i/\mu_{ji} \leq 1, \quad j \in \bar{\mathcal{J}} \quad \text{over} \quad \Lambda_i \geq 0, \quad i \in \mathcal{I}.$$

In this optimization, all constraints that are not relevant to our solution $\Lambda^*(n)$ are removed. Thus, it is not surprising that the following lemma holds.

Lemma 2. $\bar{\Lambda}^*(n) = \Lambda^*(n)$.

We prove Lemma 2 in Appendix A. A direct consequence of Lemma 2 and Proposition 1 is the following.

Lemma 3.

$$\bar{\Lambda}_i(n) = \frac{\bar{B}(n-e_i)}{\bar{B}(n)} \xrightarrow[c\to\infty]{} \Lambda_i^*(n), \qquad i \in \mathcal{I}.$$

We can now proceed to demonstrate that (3.2) holds. Let us consider the equilibrium distribution $\pi^\circ(m^\circ|n)$. We have

$$\begin{aligned}
\pi^\circ(m^\circ|n) &= \sum_{\substack{\bar{m}\in\mathbb{Z}_+^{\bar{\mathcal{J}}}:\\ (\bar{m},m^\circ)\in\mathcal{S}(n)}} \pi((\bar{m},m^\circ)|n) \\
&= \frac{1}{B(n)} \times \prod_{j\in\mathcal{J}^\circ} \left(\binom{m_j^\circ}{m_{ji}^\circ : j\in i} \prod_{i:j\in i} \mu_{ji}^{-m_{ji}^\circ} \right) \times \sum_{\substack{\bar{m}\in\mathbb{Z}_+^{\bar{\mathcal{J}}}:\\ (\bar{m},m^\circ)\in\mathcal{S}(n)}} \prod_{j\in\mathcal{J}} \left(\binom{\bar{m}_j}{\bar{m}_{ji} : j\in i} \prod_{i:j\in i} \mu_{ji}^{-\bar{m}_{ji}} \right) \\
&= \frac{1}{B(n)} \times \prod_{j\in\mathcal{J}^\circ} \left(\binom{m_j^\circ}{m_{ji}^\circ : j\in i} \prod_{i:j\in i} \mu_{ji}^{-m_{ji}^\circ} \right) \times \sum_{\bar{m}\in\bar{\mathcal{S}}(n-n^\circ)} \prod_{j\in\mathcal{J}} \left(\binom{\bar{m}_j}{\bar{m}_{ji} : j\in i} \prod_{i:j\in i} \mu_{ji}^{-\bar{m}_{ji}} \right) \\
&= \underbrace{\frac{\bar{B}(n)}{B(n)}}_{\text{(c)}} \times \underbrace{\frac{\bar{B}(n-n^\circ)}{\bar{B}(n)}}_{\text{(b)}} \times \underbrace{\prod_{j\in\mathcal{J}^\circ} \left(\binom{m_j^\circ}{m_{ji}^\circ : j\in i} \prod_{i:j\in i} \mu_{ji}^{-m_{ji}^\circ} \right)}_{\text{(a)}}. \qquad (4.3)
\end{aligned}$$

Here, $n^\circ = (n_i^\circ : i \in \mathcal{I})$ is the number of route-i customers in non-bottleneck queues, i.e. $n_i^\circ = \sum_{j\in i\cap\mathcal{J}^\circ} m_{ji}^\circ$. The third equality above follows by observing that the summation is over all the states where the number of non-bottleneck customers is $n - n^\circ$, whch then gives $B(n-n^\circ)$.

We now consider how the terms (a), (b) and (c), converge as we keep m° fixed and let n increase. Term (a) is easily dealt with as it does not depend on n. Term (b) will be shown to converge in the next proposition. Subsequently, term (c) will take a more in depth analysis.

Term (a) represents the correct expression for the unnormalized stationary distribution of our open queueing network (see (2.9)), except that we do not include the multiplicative term

$$\prod_{j\in\mathcal{J}^\circ} \prod_{i:j\in i} \Lambda_i^*(n)^{m_{ji}^\circ} = \prod_{i\in\mathcal{I}} \Lambda_i^*(n)^{n_i^\circ}.$$

As the following proposition shows, this is the limit of the term (b).

Proposition 2. *For $n^\circ \in \mathbb{Z}_+^I$, $n \in \mathbb{R}_+^I$ and $\{u_c\}_{c\in\mathbb{N}}$ some bounded sequence such that $cn + u_c \in \mathbb{Z}_+^I$*

$$\frac{B(cn+u_c-n^\circ)}{B(cn+u_c)} \xrightarrow[c\to\infty]{} \prod_{i\in\mathcal{I}} \Lambda_i^*(n)^{n_i^\circ}.$$

Proof. Proof From Proposition 1, we have that

$$\frac{B(cn+u_c-e_i)}{B(cn+u_c)} \xrightarrow[c\to\infty]{} \Lambda_i^*(n), \qquad (4.4)$$

for any $n \in \mathbb{R}_+^I$ and any bounded sequence $\{u_c\}_{c\in\mathbb{N}}$ such that $cn+u_c \in \mathbb{Z}_+^I$. Let $K = \sum_i n_i^\circ$, Let $e_{i(1)}, ..., e_{i(K)}$ be a finite sequence of unit vectors and $n(0), ..., n(K)$ be a sequence of vectors in $\mathbb{Z}_+^I$ such that

$$\begin{aligned}
&n(0) = 0, \quad n(K) = n^\circ \\
&n(k) = n(k-1) + e_{i(k)}, \quad k = 1, ..., K.
\end{aligned}$$

Applying (4.4), we have that

$$\frac{B(cn+u_c-n^\circ)}{B(cn+u_c)} = \prod_{k=0}^{K-1} \frac{B(cn+u_c-n(k)-e_{i(k+1)})}{B(cn+u_c-n(k))} \xrightarrow[c\to\infty]{} \prod_{k=1}^{K} \Lambda_{i(k)}^*(n) = \prod_{i\in\mathcal{I}} \Lambda_i^*(n)^{n_i^\circ},$$

as required. ◻

As the above proposition holds for any closed queueing network and because Lemma 3 holds, we can say that

$$\frac{\bar{B}(cn+u_c-n^\circ)}{\bar{B}(cn+u_c)} \xrightarrow[c\to\infty]{} \prod_{i\in\mathcal{I}} \Lambda_i^*(n)^{n_i^\circ} = \prod_{j\in\mathcal{J}^\circ}\prod_{i:j\in i} \Lambda_i^*(n)^{m_{ji}^\circ}. \tag{4.5}$$

This gives the convergence of term (b) in expression (4.3).

We now study term (c) in expression (4.3). We can note that this term is exactly the probability that all non-bottleneck queues are empty.

Lemma 4.

$$\pi(\{m_j=0,\forall j\in\mathcal{J}^\circ\}|n) = \frac{\bar{B}(n)}{B(n)}.$$

Proof. Proof We have the following equations:

$$\begin{aligned}
\pi(\{m_j=0,\forall j\in\mathcal{J}^\circ\}|n) &= \sum_{\substack{m\in\mathcal{S}(n):\\ m_j=0,j\in\mathcal{J}^\circ}} \frac{1}{B(n)}\prod_{j\in\mathcal{J}}\left(\binom{m_j}{m_{ji}:j\in i}\prod_{i:j\in i}\mu_{ji}^{-m_{ji}}\right)\\
&= \sum_{(0,\bar{m})\in\mathcal{S}(n)} \frac{1}{B(n)}\prod_{j\in\bar{\mathcal{J}}}\left(\binom{\bar{m}_j}{\bar{m}_{ji}:j\in i}\prod_{i:j\in i}\mu_{ji}^{-\bar{m}_{ji}}\right)\\
&= \sum_{\bar{m}\in\bar{\mathcal{S}}(n)} \frac{1}{B(n)}\prod_{j\in\bar{\mathcal{J}}}\left(\binom{\bar{m}_j}{\bar{m}_{ji}:j\in i}\prod_{i:j\in i}\mu_{ji}^{-\bar{m}_{ji}}\right) = \frac{\bar{B}(n)}{B(n)}.
\end{aligned}$$

□

Although it is difficult to directly deal with events of the form $\{m_j=0,\forall j\in\mathcal{J}^\circ\}$, we can deal with events of the form $\{m_j>0,\forall j\in\mathcal{J}'\}$ where $\mathcal{J}'\subset\mathcal{J}^\circ$.

Lemma 5. *For $\mathcal{J}'\subset\mathcal{J}^\circ$,*

$$\pi(\{m_j>0,\forall j\in\mathcal{J}'\}|cn+u_c) \xrightarrow[c\to\infty]{} \prod_{j\in\mathcal{J}'} U_j^*(n)$$

where we define

$$U_j^*(n) \stackrel{\text{def}}{=} \sum_{i:j\in i}\frac{\Lambda_i^*(n)}{\mu_{ji}}. \tag{4.6}$$

Proof. Proof To prove this lemma, we consider the explicit stationary distribution of a closed multi-class queueing network (2.1).

Recall $s=(s_j:j\in\mathcal{J})$ where $s_j=(i_j(k):k=1,...,m_j)\in\mathcal{I}^{m_j}$ keeps track of the exact position and route type of each customer within a queue. With this state representation, we can calculate the probability that the customer at the head of each queue $j\in\mathcal{J}'$ is from a specific route type $r(j)\in\mathcal{I}$. In particular, if we let $r(j)$ be the route type of the customer at the head of queue $j\in\mathcal{J}'$ and we let the vector n' give the number of customers of each route type at the heads of these queues then we can see that

$$\begin{aligned}
\pi(\{i_j(1)=r(j),j\in\mathcal{J}'\}|n) &= \sum_{\substack{s\in\mathcal{X}(n):\\ i_j(1)=r(j),j\in\mathcal{J}'}} \frac{1}{B(n)}\prod_{j\in\mathcal{J}}\prod_{i:j\in i}\mu_{ji}^{-m_{ji}}\\
&= \left[\prod_{j\in\mathcal{J}'}\frac{1}{\mu_{jr(j)}}\right]\times\sum_{s'\in\mathcal{X}(n-n')}\frac{1}{B(n)}\prod_{j\in\mathcal{J}}\prod_{i:j\in i}\mu_{ji}^{-m'_{ji}}\\
&= \left[\prod_{j\in\mathcal{J}'}\frac{1}{\mu_{jr(j)}}\right]\times\frac{B(n-n')}{B(n)}.
\end{aligned}$$

In the second inequality above, we factor out the multiplicative terms corresponding to the heads of each queue in $\mathcal{J}'$. We then notice the remaining states that must be summed over are all the states where there are $n-n'$ customers on each route. The resulting sum then gives the normalizing constant when there are $n-n'$ customers on each route.

Next, by Proposition 2, we have that

$$\pi(\{i_j(1)=r(j),j\in\mathcal{J}'\}|n) = \left[\prod_{j\in\mathcal{J}'}\frac{1}{\mu_{jr(j)}}\right]\times\frac{B(n-n')}{B(n)} \xrightarrow[c\to\infty]{} \prod_{j\in\mathcal{J}'}\frac{\Lambda_{r(j)}^*(n)}{\mu_{jr(j)}}. \tag{4.7}$$

The events $\{i_j(1)=r(j),j\in\mathcal{J}'\}$ are disjoint for different choices of $r'=(r(j):j\in\mathcal{J}')$. For a queue to be nonempty, there must be some customer at its head. Thus,

$$\bigcup_{r'\in\mathcal{I}^{J'}}\{i_j(1)=r(j),j\in\mathcal{J}'\} = \{m_j>0,j\in\mathcal{J}\},$$

and consequently using (4.7), we have

$$\begin{aligned}\pi(\{m_j>0, j\in\mathcal{J}\}|n) &= \sum_{r'\in\mathcal{I}^{J'}} \pi(\{i_j(1)=r(j), j\in\mathcal{J}'\}|n)\\ &= \sum_{r'\in\mathcal{I}^{J'}} \left[\prod_{j\in\mathcal{J}'}\frac{1}{\mu_{jr(j)}}\right]\times\frac{B(n-n')}{B(n)}\\ &\xrightarrow[c\to\infty]{} \sum_{r'\in\mathcal{I}^{J'}}\prod_{j\in\mathcal{J}'}\left(\frac{\Lambda^*_{r(j)}(n)}{\mu_{jr(j)}}\right)=\prod_{j\in\mathcal{J}'}\left(\sum_{r\in\mathcal{I}}\frac{\Lambda^*_r(n)}{\mu_{jr}}\right)=\prod_{j\in\mathcal{J}'}U^*_j(n)\end{aligned}$$

as required. □

Now we are in the position to prove the convergence of term (c) in expression (4.3).

Proposition 3.

$$\frac{\bar{B}(cn+u_c)}{B(cn+u_c)}\xrightarrow[c\to\infty]{}\prod_{j\in\mathcal{J}^\circ}\left(1-U^*_j(n)\right)$$

Proof. Proof In the following expression, we use Lemma 4; we apply the Inclusion-Exclusion Principle (4.9); we apply Lemma 5 (4.10); and then we notice the resulting summation is $\prod_{j\in\mathcal{J}^\circ}\left(1-U^*_j(n)\right)$ expanded:

$$\begin{aligned}\frac{\bar{B}(cn+u_c)}{B(cn+u_c)} =&\pi(\{m_j=0,\forall j\in\mathcal{J}^\circ\}|cn+u_c) &(4.8)\\ =&1-\pi\Big(\bigcup_{j\in\mathcal{J}^\circ}\{m_j>0\}\Big|cn+u_c\Big)\\ =&\sum_{k=0}^{J^\circ}\sum_{\substack{j_1,\dots,j_k\in\mathcal{J}^\circ:\\ j_1<\dots<j_k}}(-1)^k\pi(\{m_j>0, j=j_1,\dots,j_k\}|cn+u_c) &(4.9)\\ \xrightarrow[c\to\infty]{}&\sum_{k=0}^{J^\circ}\sum_{\substack{j_1,\dots,j_k\in\mathcal{J}^\circ:\\ j_1<\dots<j_k}}(-1)^k\prod_{j=j_1,\dots,j_k}U^*_j(n)=\prod_{j\in\mathcal{J}^\circ}\left(1-U^*_j(n)\right). &(4.10)\end{aligned}$$

In expression (4.9), the $k=0$ summand is understood to equal 1. □

We have now discovered the limiting behavior of terms (a), (b), and (c), and we can prove Theorem 1.

Proof. Proof of Theorem 1 We apply to the equality (4.3) the results (4.5) (which we derived from Proposition 3) and Proposition 2:

$$\begin{aligned}\pi^\circ(m^\circ|cn+u_c)=&\frac{\bar{B}(cn+u_c)}{B(cn+u_c)}\times\frac{\bar{B}(cn+u_c-n^\circ)}{\bar{B}(cn+u_c)}\times\prod_{j\in\mathcal{J}^\circ}\left(\binom{m^\circ_j}{m^\circ_{ji}:j\in i}\prod_{i:j\in i}\frac{1}{\mu_{ji}^{m^\circ_{ji}}}\right)\\ \xrightarrow[c\to\infty]{}&\prod_{j\in\mathcal{J}^\circ}\left(1-U^*_j(n)\right)\times\prod_{j\in\mathcal{J}^\circ}\prod_{i:j\in i}\Lambda^*_i(n)^{m^\circ_{ji}}\times\prod_{j\in\mathcal{J}^\circ}\left(\binom{m^\circ_j}{m^\circ_{ji}:j\in i}\prod_{i:j\in i}\left(\frac{1}{\mu_{ji}}\right)^{m^\circ_{ji}}\right)\\ =&\prod_{j\in\mathcal{J}^\circ}\left(1-U^*_j(n)\right)\times\prod_{j\in\mathcal{J}^\circ}\left(\binom{m^\circ_j}{m^\circ_{ji}:j\in i}\prod_{i:j\in i}\left(\frac{\Lambda_i(n)}{\mu_{ji}}\right)^{m^\circ_{ji}}\right)\\ =&\pi^\circ_{\Lambda^*(n)}(m^\circ).\end{aligned}$$

This proves that our network's non-bottleneck queues become independent. □

We have now shown, for our stationary closed queueing network, convergence in distribution of non-bottleneck queues to a stationary open queueing network. We now wish to show convergence of queue sizes in expectation, Theorem 2.

Building on the above results, we now develop a proof for Theorem 2. Once again, we gain the result by considering a modified version of our original queueing network. Recalling the normalizing constant $B(n)$ (2.3), let $B^{+j}(n)$ be the normalizing constant of the closed queueing network that is obtained by adding a replica of queue j. This replica is added immediately after queue j. Customers leaving j will immediately join this replica queue, and then, after service at the replica queue, customers will continue on their route. We can now express mean queue lengths in terms of these normalizing constants.

Lemma 6.

$$EM^\circ_{ji}(n)=\frac{1}{\mu_{ji}}\frac{B^{+j}(n-e_i)}{B(n)}.$$

Proof. Proof For all $j' \in \mathcal{J}, i' \in \mathcal{I}$, we have

$$EM^{\circ}_{j'i'}(n) \stackrel{\text{def}}{=} \frac{1}{B(n)} \sum_{m \in \mathcal{S}(n): m_{j'i'}>0} m_{j'i'} \prod_{j \in \mathcal{J}} m_j! \prod_{i:j\in i} \frac{\mu_{ji}^{-m_{ji}}}{m_{ji}!} \tag{4.11}$$

$$= \frac{\mu_{j'i'}^{-1}}{B(n)} \sum_{m \in \mathcal{S}(n): m_{j'i'}>0} \prod_{\substack{j\in\mathcal{J}:\\ j\neq j'}} m_j! \prod_{i:j\in i} \frac{\mu_{ji}^{-m_{ji}}}{m_{ji}!} \times m_{j'}! \prod_{i:j'\in i, i\neq i'} \frac{\mu_{j'i}^{-m_{j'i}}}{m_{j'i}!} \times \frac{\mu_{j'i'}^{-m_{j'i'}+1}}{(m_{j'i'}-1)!} \tag{4.12}$$

$$= \frac{\mu_{j'i'}^{-1}}{B(n)} \sum_{m \in \mathcal{S}(n-e_{i'})} \prod_{\substack{j\in\mathcal{J}:\\ j\neq j'}} m_j! \prod_{i:j\in i} \frac{\mu_{ji}^{-m_{ji}}}{m_{ji}!} \times (m_{j'}+1)! \prod_{i:j'\in i} \frac{\mu_{j'i}^{-m_{j'i}}}{m_{j'i}!} \tag{4.13}$$

$$= \frac{\mu_{j'i'}^{-1}}{B(n)} \sum_{m \in \mathcal{S}(n-e_{i'})} \prod_{\substack{j\in\mathcal{J}:\\ j\neq j'}} m_j! \prod_{i:j\in i} \frac{\mu_{ji}^{-m_{ji}}}{m_{ji}!} \times \sum_{k=0}^{m_{j'}} \binom{m_{j'}}{m_{j'i} : i \ni j} \prod_{i:j'\in i} \mu_{j'i}^{-m_{j'i}} . \tag{4.14}$$

Using the multinomial Vandermonde convolution we can rewrite the last sum in (4.14) in the following way

$$\sum_{k=0}^{m_{j'}} \binom{m_{j'}}{m_{j'i} : i \ni j} \prod_{i:j'\in i} \mu_{j'i}^{-m_{j'i}} \tag{4.15}$$

$$= \sum_{k=0}^{m_{j'}} \sum_{\substack{\tilde{m}_1,\tilde{m}_2 \in \mathbb{Z}_+^I : \sum_i \tilde{m}_{1i}=k,\\ \tilde{m}_{1i}+\tilde{m}_{2i}=m_{j'i},\forall i}} \binom{k}{\tilde{m}_{1i} : i \ni j'} \binom{m_{j'}-k}{m_{j'i}-\tilde{m}_{1i} : i \ni j'} \prod_{i:j'\in i} \mu_{j'i}^{-m_{j'i}} \tag{4.16}$$

$$= \sum_{\substack{\tilde{m}_1,\tilde{m}_2 \in \mathbb{Z}_+^I :\\ \tilde{m}_{1i}+\tilde{m}_{2i}=m_{j'i},\forall i}} \binom{\sum_i \tilde{m}_{1i}}{\tilde{m}_{1i} : i \ni j'} \prod_{i:j'\in i} \mu_{j'i}^{-\tilde{m}_{1i}} \binom{\sum_i \tilde{m}_{2i}}{\tilde{m}_{2i} : i \ni j'} \prod_{i:j'\in i} \mu_{j'i}^{-\tilde{m}_{2i}} . \tag{4.17}$$

We can interpret the above identity as adding an extra queue to our network and hence it follows that

$$EM^{\circ}_{j'i'}(n) = \frac{\mu_{j'i'}^{-1}}{B(n)} \sum_{m \in \mathcal{S}^{+j'}(n-e_{i'})} \prod_{j \in \mathcal{J}^{+j'}} m_j! \prod_{i:j\in i} \frac{\mu_{ji}^{-m_{ji}}}{m_{ji}!} = \frac{\mu_{j'i'}^{-1}}{B(n)} B^{+j'}(n-e_{i'}) . \tag{4.18}$$

In the above expression, $\mathcal{S}^{+j'}(n-e_{i'})$ and $\mathcal{J}^{+j'}$ are, respectively, the state space and set of queues obtained when adding a replica of queue j'. □

With Lemma 6, we can prove Theorem 2.

Proof. Proof of Theorem 2 By assumption, $j \in \mathcal{J}^{\circ}$ – it is a non-bottleneck queue. Notice that the replica queue added for B^{+j} corresponds to a repeated constraint in optimization (3.3), which by Theorem 1 implies that the replica queue must also be a non-bottleneck queue.

By Lemma 6 and the definition of $\bar{B}(n)$ given in (4.1), the following equation holds

$$\lim_{c\to\infty} EM^{\circ}_{ji}(cn+u_c) = \frac{1}{\mu_{ji}} \lim_{c\to\infty} \frac{\bar{B}(cn+u_c)}{B(cn+u_c)} \frac{\bar{B}(cn-e_i+u_c)}{\bar{B}(cn+u_c)} \frac{B^{+j}(cn-e_i+u_c)}{\bar{B}(cn-e_i+u_c)} \tag{4.19}$$

Now, using Lemma 3,

$$\lim_{c\to\infty} \frac{\bar{B}(cn-e_i+u_c)}{\bar{B}(cn+u_c)} = \Lambda_i^*(n) \tag{4.20}$$

Since the replica queue j is not a bottleneck, we can apply Proposition 3 twice to find that

$$\lim_{c\to\infty} \frac{\bar{B}(cn+u_c)}{B(cn+u_c)} \frac{B^{+j}(cn-e_i+u_c)}{\bar{B}(cn-e_i+u_c)} = \frac{1}{1-U_j^*(n)}, \tag{4.21}$$

where we recall that $U_j^*(n)$ is defined in (4.6). Substituting (4.20)-(4.21) in (4.19), we get

$$\lim_{c\to\infty} EM^{\circ}_{ji}(cn+u_c) = \frac{\Lambda_i^*(n)/\mu_{ji}}{1-U_j^*(n)}. \tag{4.22}$$

For a non-bottleneck queue, $j \in \mathcal{J}^{\circ}$, $U_j^*(n) = \sum_{i:j\in i} \Lambda_i^*(n)/\mu_{ji} < 1$, with some standard calculations on one can verify that

$$EM^{\circ}_{\Lambda^*(n),ji} = \frac{\Lambda_i^*(n)/\mu_{ji}}{1-U_j^*(n)}, \tag{4.23}$$

for all $i \in \mathcal{I}$ and $j \in \mathcal{J}^{\circ}$. Thus together (4.23) and (4.22) imply the required result

$$\lim_{c\to\infty} EM^{\circ}_{ji}(cn+u_c) = EM^{\circ}_{\Lambda^*(n),ji}$$

for all $i \in \mathcal{I}$ and $j \in \mathcal{J}^{\circ}$. □

4.2 Proof of fluid limit.

In this section, we prove Theorem 3. We first introduce and recall some notation before proceeding with a proof.

Let $\mathcal{I}(j) \stackrel{\text{def}}{=} \{i \in \mathcal{I} : j \in i\}$ and $I(j) \stackrel{\text{def}}{=} |\mathcal{I}(j)|$. Assuming that set $\mathcal{I}(j)$ is ordered, we also denote by $i_j(l)$, with $l = 1, \dots, I(j)$, the function that enumerates its elements. We define for each $j \in \mathcal{J}$, an independent Poisson marked point process N_j with intensity $\mu_j \, dt \otimes du$ on $\mathbb{R} \times [0,1]$, where $\mu_j = \max\{ \mu_{ji} : i \in \mathcal{I}(j)\}$, and the function $\chi_j(m_j, u)$ on $\mathbb{N}^{I(j)} \times (0,1)$ in the following way:

$$\chi_j(m_j,u) = \begin{cases} 0 & \text{if } m_j = 0 \\ i_j(l) & \text{if } \sum_{h=0}^{l-1} m_{jh}\,\mu_{jh} \le m_j\,\mu_j \times u \le \sum_{h=0}^{l} m_{jh}\,\mu_{jh} \\ 0 & \text{if } \sum_{h=0}^{I(j)} m_{jh}\,\mu_{jh} \le m_j\,\mu_j \times u \le m_j\,\mu_j. \end{cases}$$

Therefore, the network process can be written as

$$dM_{ji}(t) = 1\{\chi_{j'}(M_{j'}(t-), U_{N_{j'}(t)}) = i\}\, dN_{j'}(t) - 1\{\chi_j(M_j(t-), U_{N_j(t)}) = i\}\, dN_j(t)$$

where j' denotes the queue just before j on route i, and the uniform independent random variables U_k are generated by the second coordinate of the marked point process. In the following we will use also the notations $\hat{j}$ and $\hat{j}'$ to denote the first queue on route i before the queues j and j' respectively that are non empty. In integral form, we have

$$M_{ji}(t) = M_{ji}(0) + \mathcal{L}_{j'i}(t) - \mathcal{L}_{ji}(t) + L_{j'i}(t) - L_{ji}(t)$$

where $\mathcal{L}_{ji}(t)$ is the martingale

$$\mathcal{L}_{ji}(t) = \int_0^t \int_0^1 1\{\chi_j(M_j(s-),u) = i\}[dN_j(s,u) - \mu_j\, du\, ds],$$

and the process $L_{ij}(t)$ is given by

$$L_{ji}(t) = \int_0^t \int_0^1 \mu_j\, 1\{\chi_j(M_j(s-),u) = i\}\, du\, ds.$$

From the equation above, noticing that $\int_0^1 1\{\chi_j(M_j(s,\, u) = i\}\, du = (\mu_{ji}\, M_{ji}(s))/(\mu_j\, M_j(s))$ and with the convention that $0/0 = 0$, we get that

$$\frac{1}{\mu_{ji}} L_{ji}(t) = \int_0^t \frac{\mu_j}{\mu_{ji}} \int_0^1 1\{\chi_j(M_j(s-),u) = i\}\, du\, ds = \int_0^t \frac{M_{ji}(s-)}{M_j(s-)}\, ds.$$

Summing over $i : j \in i$ we have that for $t' \le t''$

$$\sum_{i:j\in i} \frac{1}{\mu_{ji}} \left(L_{ji}(t'') - L_{ji}(t')\right) = \int_{t'}^{t''} \sum_{i:j\in i} \frac{M_{ji}(s-)}{M_j(s-)}\, ds = \int_{t'}^{t''} \sum_{i:j\in i} 1\{M_j(s-) \neq 0\}\, ds \le t'' - t', \tag{4.24}$$

which gives the Lipschitz condition for the process $L_{ji}(t)$.

We define the scaled process $\bar{M}(n\,c,\, t) = c^{-1}\, M(n\,c,\, c\,t)$, such that, having $M_{ji}(0) = c\,n_{ji}$, it has the ji-component given by

$$\bar{M}_{ji}(n\,c,\, t) = n_{ji} + \frac{\mathcal{L}_{j'i}(ct) - \mathcal{L}_{ji}(ct)}{c} + \frac{L_{j'i}(ct) - L_{ji}(ct)}{c}.$$

Using the same steps as before, we can rewrite the process $c^{-1} L_{ji}(ct)$ in the following way

$$\begin{aligned} \frac{L_{ji}(ct)}{c} &= \frac{1}{c} \int_0^{c\,t} \int_0^1 \mu_j\, 1\{\chi_j(M_j(s),u) = i\}\, du\, ds \\ &= \int_0^t \int_0^1 \mu_j\, 1\{\chi_j(M_j(c\,s),\, u) = i\}\, du\, ds = \int_0^t \frac{M_{ji}(c\,s)}{M_j(c\,s)}\, \mu_{ji}\, ds\ . \end{aligned}$$

Proposition 4. *Given $n \in \mathbb{N}^{J\times I}$, the processes $\{(\bar{M}(n\,c,\, t), t > 0)\}_{c>0}$, with $M(n\,c,\, 0) = n\,c$, are tight.*

Proof. Proof By the triangular inequalities for metrics, to prove tightness of the multidimensional process $\{(\bar{M}(n\,c,\, t), t > 0)\}_{c>0}$, it is enough to show that any of its coordinate processes are tight. By Theorem C.9 in Robert [36], this can be done by showing that: given i, j, for any fixed $T, \eta > 0$, there exists $\delta > 0$ such that

$$\Pr\{\omega_{\bar{M}_{ji}(n\,c,\,\cdot)}(\delta) > \eta\} < \epsilon \tag{4.25}$$

where, for a given function f, the modulus of continuity on $[0,T]$ is defined as

$$\omega_f(\delta) = \sup\{|f(t) - f(s)| : s,t \le T, |t-s| < \delta\}.$$

Since $\bar{M}_{ji}(n\,c,\,t) = n + c^{-1}\,\Delta\mathcal{L}_{ji}(c\,t) + c^{-1}\,\Delta L_{ji}(c\,t) - (\mathcal{L}_{ji}(c\,t))$ with $\Delta\mathcal{L}_{ji}(t) = \mathcal{L}_{j'i}(t) - \mathcal{L}_{ji}(t)$ and $\Delta L_{ji}(t) = L_{j'i}(t) - L_{ji}(t)$, it is enough to prove that relation (4.25) is valid separately for the processes $c^{-1}\Delta\mathcal{L}_{ji}(c\,t)$ and $c^{-1}\Delta L_{ji}(c\,t)$. We have, for $T > 0$ and $\eta > 0$,

$$
\begin{aligned}
&\Pr\Big\{\sup_{s,t\le T;|t-s|<\delta}\Big|\frac{\Delta L_{ij}(c\,t)}{c} - \frac{\Delta L_{ij}(c\,s)}{c}\Big| > \eta\Big\} \\
&= \Pr\Big\{\sup_{s,t\le T;|t-s|<\delta}\Big|\int_s^t \left(\frac{M_{j'i}(c\,u)}{M_{j'}(c\,u)}\,\mu_{j'i} - \frac{M_{ji}(c\,u)}{M_j(c\,u)}\,\mu_{ji}\right)\,du\Big| > \eta\Big\} \\
&\le \Pr\Big\{\sup_{s,t\le T;|t-s|<\delta}\int_s^t \left|\frac{M_{j'i}(c\,u)}{M_{j'}(c\,u)}\,\mu_{j'i} - \frac{M_{ji}(c\,u)}{M_j(c\,u)}\,\mu_{ji}\right|\,du > \eta\Big\} \\
&\le \Pr\Big\{\sup_{s,t\le T;|t-s|<\delta} 2|t-s|\mu > \eta\Big\} = 0
\end{aligned}
\tag{4.26}
$$

with $\delta < \eta/(2\mu)$ and where $\mu = \max\{i, j : \mu_{ji}\}$.

For the martingale $c^{-1}\,\Delta\mathcal{L}_{ji}(c\,t)$, we have that

$$
\Pr\Big\{\sup_{s,t\le T;|t-s|<\delta}\Big|\frac{\Delta\mathcal{L}_{ji}(c\,t)}{c} - \frac{\Delta\mathcal{L}_{ji}(c\,s)}{c}\Big| > \eta\Big\} = \Pr\Big\{\sup_{t\le T}\frac{\Delta\mathcal{L}_{ji}(c\,t)}{c} > \frac{\eta}{2}\Big\} \le \frac{4}{c^2\eta^2}\mathbb{E}\left[(\Delta\mathcal{L}_{ji}(c\,t))^2\right]
$$

where the last inequality follows by applying the Doob's inequality. Having that $\mathbb{E}\left[(\Delta\mathcal{L}_{ji}(c\,T))^2\right] \le 2\mu c T$ we have that for $c > 8\mu T/(\epsilon\eta^2)$ the probability is bounded above by ϵ, as required. □

The tightness property ensures the relative compactness, therefore from every sequence $\{(\bar{M}(n\,c,\,t), t > 0)\}_{c>0}$, with $M(n\,c,\,0) = n\,c$ and n fixed, it is possible to extract a convergent subsequence. The following proposition ensures that any limit process will be given by a fluid solution.

Proposition 5. *Assume that a sequence of processes $\{(\bar{M}(n\,c,\,t), t > 0)\}_{c>0}$ converges to a limit process $m(t)$ as $c \to \infty$. Then, $m(t)$ is almost surely continuous and it is a fluid solution as defined in Definition 2.*

Proof. Proof Using the Skorohod's Representation theorem, see Robert [36, Theorem C.8], we can assume that all the elements of the sequence are random processes defined on the same probability space with probability $\mathbb{P}$ and the convergence is $\mathbb{P}$-a.s.

Since we have that for any c

$$
\bar{M}_{ji}(n\,c,\,t) = n_{ji} + \frac{\Delta\mathcal{L}_{ji}(c\,t)}{c} + \frac{L_{j'i}(c\,t)}{c} - \frac{L_{ji}(c\,t)}{c},
$$

passing to the limit and using bounded convergence for the integral we get

$$
\bar{m}_{ji}(t) = n_{ji} + L_{j'i}(t) - L_{ji}(t)
$$

where

$$
L_{ji}(t) = \lim_{c\to\infty}\int_0^t \frac{M_{ji}(c\,u)}{M_j(c\,u)}\,\mu_{ji}\,du,
$$

which exists by (4.26). In particular, if $m_j(t) > 0$ for some t, it will be positive in a neighborhood $B(t)$ of t by continuity. It follows that for $t' < t''$ and $t', t'' \in B(t)$,

$$
L_{ji}(t'') - L_{ji}(t') = \lim_{c\to\infty}\int_{t'}^{t''} \frac{M_{ji}(c\,u)}{M_j(c\,u)}\,\mu_{ji}\,du = \int_{t'}^{t''} \lim_{c\to\infty}\frac{M_{ji}(c\,u)}{M_j(c\,u)}\,\mu_{ji}\,du = \int_{t'}^{t''} \frac{m_{ji}(u)}{m_j(u)}\,\mu_{ji}\,du,
$$

where in the second equality we have used the bounded convergence theorem, which implies equation (2.10d), i.e.

$$
\Lambda_{ji}(t) = \frac{dL_{ji}(t)}{dt} = \mu_{ji}\,\frac{m_{ji}(t)}{m_j(t)} \quad \text{as } m_j(t) > 0.
$$

The additional conditions satisfied by $L_{ji}(t)$ easily follow as equivalent property holds for the approximating processes $\{c^{-1}\,L_{ji}(c\,t), c > 0\}$, in particular the Lipschitz condition follows by (4.24). □

4.3 Proof of convergence of bottleneck queues.

We now focus on proving Theorem 4. For technical reasons that we will explain shortly, our proof is quite involved and requires a number of lemmas. For a solution $(m(t), L(t))$ of the fluid model (2.10), let $\Lambda_{ji}(t)$ be the derivative of $L_{ji}(t)$, when it exists. Let also l_{ji} be the queue before queue j on route i and x_{ji} be the next queue after queue j on route i.

We recall the function $\beta(m(t))$:

$$
\beta(m(t)) \stackrel{\text{def}}{=} \sum_{j\in\mathcal{J}}\ \sum_{\substack{i:j\in i\\ m_{ji}(t)>0}} m_{ji}(t)\log\frac{m_{ji}(t)\mu_{ji}}{m_j(t)}.
\tag{4.27}
$$

The terms $m_{ji}(t)$ are Lipschitz and thus are differentiable for almost every t. So, if $m_{ji}(t) > 0$ then we can differentiate the ij^{th} summand of (4.27). However, if $m_{ji}(t) = 0$, taking a derivative become a significantly more technical issue. The following proposition ensures we can differentiate the summands of $\beta(m(t))$ and the subsequent lemma ensures summands have zero derivative when $m_{ji}(t) = 0$.

Proposition 6. *For each i and $j \in i$ and for any time interval $[t_0, t]$ with $t > t_0 > 0$, there exists a constant $D > 0$ such that for any t_1, $t_2 \in [t_0, t]$*

$$|\beta(m(t_2)) - \beta(m(t_1))| \le D|t_2 - t_1| \ . \tag{4.28}$$

Moreover, the same locally-Lipschitz condition applies to each summand of $\beta(m(t))$, (4.27).

Lemma 7. *For functions $f : \mathbb{R}_+ \to \mathbb{R}_+$ and $p : \mathbb{R}_+ \to [0,1]$, if $t > 0$ is such that the derivative of f exists at t, the derivative of $f \log p$ exists at t and $f(t) = 0$ then*

$$\frac{df}{dt} = 0 \qquad \text{and} \qquad \frac{df \log p}{dt} = 0.$$

Proposition 6 is based on Proposition 4.2 of Bramson [7]. Both Proposition 6 and Lemma 7 are proven in Appendix B. It will also be useful to have the following lemma, which is proven in Appendix B.

Lemma 8. *For almost every t, $0 < \Lambda_{ji}(t) \le \mu_{max}$, where $\mu_{max} = \max\{\mu_{ji} : i \in \mathcal{I}, j \in i\}$.*

Recalling that x_{ji} is the next queue on route i after j and that l_{ji} is the queue before queue j on route i, we can now prove the following proposition.

Proposition 7. *For almost every t,*

$$\frac{d\beta(m(t))}{dt} = -\sum_{i\in\mathcal{I}}\sum_{j\in i}\Lambda_{ji}(t)\log\frac{\Lambda_{ji}(t)}{\Lambda_{x_{ji}i}(t)}. \tag{4.29}$$

Proof. Proof Our processes in Proposition 6 are absolutely continuous and thus almost everywhere differentiable. So for almost every t, we can differentiate the terms $L_{ji}(t)$, $m_{ij}(t)$, $m_j(t)$, $m_{ij}(t)\log(m_{ji}(t)\mu_{ji}/m_j(t))$, and $\beta(m(t))$. Differentiating, we obtain

$$\frac{d\beta(m(t))}{dt} = \sum_{j\in\mathcal{J}}\sum_{\substack{i:j\in i\\ m_{ji}(t)>0}}\Big(\frac{dm_{ji}\log m_{ji}}{dt} - \frac{dm_{ji}}{dt}\log\mu_{ji}\Big) - \sum_{\substack{j\in\mathcal{J}:\\ m_j(t)>0}}\frac{dm_j\log m_j}{dt} \tag{4.30a}$$

$$= \sum_{j\in\mathcal{J}}\sum_{\substack{i:j\in i\\ m_{ji}(t)>0}}\Big(\frac{dm_{ji}}{dt}\log m_{ji} + \frac{dm_{ji}}{dt} - \frac{dm_{ji}}{dt}\log\mu_{ji}\Big) - \sum_{\substack{j\in\mathcal{J}:\\ m_j(t)>0}}\Big(\frac{dm_j}{dt}\log m_j + \frac{dm_j}{dt}\Big) \tag{4.30b}$$

$$= \sum_{j\in\mathcal{J}}\sum_{\substack{i:j\in i\\ m_{ji}(t)>0}}\frac{dm_{ji}}{dt}\log\frac{m_{ji}(t)\mu_{ji}}{m_j(t)} \tag{4.30c}$$

$$= \sum_{i\in\mathcal{I}}\sum_{j\in i}\Big(\Lambda_{l_{ji}i}(t) - \Lambda_{ji}(t)\Big)\log\Lambda_{ji}(t) \tag{4.30d}$$

$$= \sum_{i\in\mathcal{I}}\sum_{j\in i}\Lambda_{ji}(t)\Big(\log\Lambda_{x_{ji}i}(t) - \log\Lambda_{ji}(t)\Big) \tag{4.30e}$$

$$= -\sum_{i\in\mathcal{I}}\sum_{j\in i}\Lambda_{ji}(t)\log\frac{\Lambda_{ji}(t)}{\Lambda_{x_{ji}i}(t)}. \tag{4.30f}$$

For the above sequence of equalities, in equality (4.30a), we restrict our attention to summands with $m_{ji}(t) > 0$ by applying Lemma 7 and Proposition 6. Equality (4.30c) holds by observing that $\sum_{i:j\in i} m_{ji} = m_j$ and canceling terms. For equality (4.30d), we know by our fluid model assumption (2.10a) that $m'_{ji}(t) = \Lambda_{l_{ji}i}(t) - \Lambda_{ji}(t)$. In addition, we note that if $m_{ji} > 0$ then, using (2.10d), $\log\frac{m_{ji}\mu_{ji}}{m_j} = \log\Lambda_{ji}(t)$ and if $m_{ji}(t) = 0$ then $0 = m'_{ji}(t) = \Lambda_{l_{ji}i}(t) - \Lambda_{ji}(t)$ and $\Lambda_{ji}(t) > 0$ (by Lemma 8). Thus, we may reintroduce the $m_{ji}(t) = 0$ terms in our summation. In equality (4.30d), for each route, we re-interpolate the first term in our summation. □

The next lemma, found in Cover and Thomas [14], will be of key importance in bounding our Lyapunov function.

Lemma 9 (Pinsker's Inequality)**.** *For the relative entropy between two discrete probability distributions $p = (p_j)_j$ and $q = (q_j)_j$ with the same support:*

$$D(p||q) = \sum_j p_j \log\frac{p_j}{q_j},$$

the following inequality holds

$$\sqrt{D(p||q)} \ge \sum_j |p_j - q_j|.$$

Applying Pinsker's inequality to Proposition 7 gives

Lemma 10. *For almost every t,*

$$\frac{d\beta(m(t))}{dt} \leq -\sum_{i\in\mathcal{I}} \frac{1}{\mu_{max}|\mathcal{J}|} \sum_{j\in i} \Big(\Lambda_{ji}(t) - \Lambda_{x_{ji}i}(t)\Big)^2, \tag{4.31}$$

where $|\mathcal{J}|$ is the size of set $\mathcal{J}$ and we recall that $\mu_{max} = \max\{\mu_{ji} : i \in \mathcal{I}, j \in i\}$.

Proof. Proof Let t be a time for which Proposition 7 holds and let $\Lambda_i^{\Sigma}(t) = \sum_{j\in i} \Lambda_{ji}(t)$, for $i \in \mathcal{I}$. For each i, let $p_j = \Lambda_{ji}(t)/\Lambda_i^{\Sigma}(t)$ and $q_j = \Lambda_{x_{ji}i}(t)/\Lambda_i^{\Sigma}(t)$. By Lemma 8, p and q both have the same support. For each i, we applying Pinsker's Lemma

$$\sum_{j\in i} \frac{\Lambda_{ji}(t)}{\Lambda_i^{\Sigma}(t)} \log \frac{\Lambda_{ji}(t)}{\Lambda_{x_{ji}i}(t)} = \sum_{j\in i} p_j \log \frac{p_j}{q_j} \geq \Big(\sum_{j\in i} |p_j - q_j|\Big)^2 \geq \sum_{j\in i} |p_j - q_j|^2 = \frac{1}{(L_i^{\Sigma}(t))^2} \sum_{j\in i} \Big(\Lambda_{ji}(t) - \Lambda_{x_{ji}i}(t)\Big)^2$$

Multiplying the left and right of this inequality by $-L_i^{\Sigma}(t)$, summing over $i \in \mathcal{I}$ gives

$$\frac{d\beta(m(t))}{dt} = -\sum_{i\in\mathcal{I}} \sum_{j\in i} \Lambda_{ji}(t) \log \frac{\Lambda_{ji}(t)}{\Lambda_{x_{ji}i}(t)} \leq -\sum_{i\in\mathcal{I}} \frac{1}{L_i^{\Sigma}(t)} \sum_{j\in i} \Big(\Lambda_{ji}(t) - \Lambda_{x_{ji}i}(t)\Big)^2.$$

Recall that from Lemma 8 that $\Lambda_{ji}(t) \leq \mu_{max}$ thus $\Lambda_i^{\Sigma}(t) \leq |\mathcal{J}|\mu_{max}$. Applying this bound to $\Lambda_i^{\Sigma}(t)$ the above equation gives the required result (4.31). □

We define m^* to be a solution to the optimization problem

$$\text{minimize} \quad \beta(m) \quad \text{subject to} \quad \sum_{j\in i} m_{ij} = n_i, \quad i \in \mathcal{I} \quad \text{over} \quad m_{ji} \geq 0, \quad i \in \mathcal{I},\ j \in i. \tag{4.32}$$

As we discussed, we expect the path of the $m(t)$ to converge to the optimal value of the optimization. To conduct further analysis, we characterize the dual of this problem.

Lemma 11. *The dual of the optimization* (4.32) *is*

$$\textit{maximize} \quad \sum_{i\in\mathcal{I}} n_i \log \Lambda_i \quad \textit{subject to} \quad \sum_{i:j\in i} \frac{\Lambda_i}{\mu_{ji}} \leq 1 \quad \textit{over} \quad \Lambda_i \geq 0, \ \ i \in \mathcal{I}.$$

Proof. Proof Taking Lagrange multipliers $\lambda \in \mathbb{R}^I$, its Lagrangian function is,

$$\begin{aligned} L(m,\lambda) &= \sum_{j\in\mathcal{J}:m_j>0} \sum_{i\in\mathcal{I}} m_{ji} \log \frac{m_{ji}\mu_{ji}}{m_j} + \sum_{i\in\mathcal{I}} \lambda_i \Big(n_i - \sum_{j:j\in i} m_{ji}\Big) \\ &= \sum_{j\in\mathcal{J}:m_j>0} \sum_{i\in\mathcal{I}} m_{ji} \log \frac{m_{ji}\mu_{ji}}{m_j e^{\lambda_i}} + \sum_{i\in\mathcal{I}} \lambda_i n_i \\ &= \sum_{j\in\mathcal{J}:m_j>0} m_j D(p^{(j)}||q^{(j)}) - \sum_{j\in\mathcal{J}:m_j>0} m_j \log\Big(\sum_{i:j\in i} e^{\lambda_i}\mu_{ji}^{-1}\Big) + \sum_{i\in\mathcal{I}} \lambda_i n_i. \end{aligned}$$

In the last inequality above, we let $p^{(j)} = (m_{ji}/m_j : i \ni j)$ and $q^{(j)} = (e^{\lambda_i}\mu_{ji}^{-1}/\sum_r e^{\lambda_r}\mu_{jr}^{-1} : i \ni j)$. Recalling our Remark 3 on relative entropies, this Lagrangian is minimized by taking $p^{(j)} = q^{(j)}$ for each $j \in \mathcal{J}$, where $D(p^{(j)}||q^{(j)}) = 0$, and then by minimizing over m_j. In particular, we get the Lagrange dual function

$$\min_{m\in\mathbb{R}_+^K} L(m,\lambda) = \begin{cases} \sum_{i:n_i>0} n_i\lambda_i & \text{if} \quad \sum_{i:j\in i} \frac{e^{\lambda_i}}{\mu_{ji}} \leq 1, \ \ \forall j \in \mathcal{J}, \\ -\infty & \text{otherwise.} \end{cases}$$

Thus, we find dual

$$\text{maximize} \quad \sum_{i:n_i>0} n_i\lambda_i \quad \text{subject to} \quad \sum_{i:j\in i} \frac{e^{\lambda_i}}{\mu_{ji}} \leq 1 \quad \text{over} \quad \lambda \in \mathbb{R}^I.$$

Substituting $\Lambda_i = e^{\lambda_i}$ gives the required result

$$\text{maximize} \quad \sum_{i:n_i>0} n_i \log \Lambda_i \quad \text{subject to} \quad \sum_{i:j\in i} \frac{\Lambda_i}{\mu_{ji}} \leq 1, \quad \forall j \in \mathcal{J} \quad \text{over} \quad \Lambda \in \mathbb{R}_+^I.$$

□

Lemma 12. *If, for some $\epsilon > 0$, $\beta(m(t)) \geq \beta(m^*) + \epsilon$ then there exists $\delta > 0$, $i \in \mathcal{I}$ and $j \in i$ such that*

$$|\Lambda_{ji}(t) - \Lambda_{x_{ji}i}(t)| \geq \delta.$$

Proof. Proof We develop a proof by contradiction. If this result were not true, as the set of queue sizes is compact, we could construct a sequence of times t^k, $k = 1, 2, 3, ...$ such that $\beta(m(t^k)) \geq \beta(m^*) + \epsilon$ and $m(t^k) \to \tilde{m}$, as $k \to \infty$ and

$$\sum_{i \in \mathcal{I}} \sum_{j \in i} \left| \Lambda_{ji}(t^k) - \Lambda_{ji}(t^k) \right| \xrightarrow[k \to \infty]{} 0.$$

For each queue $j \in \mathcal{I}$ with $\tilde{m}_j > 0$, let j_i^+ be the next non-empty queue on route i i.e. $\tilde{m}_{j_i^+ i} > 0$. We can say

$$\Lambda_{ji}(t^k) \xrightarrow[k \to \infty]{} \frac{\tilde{m}_{ji} \mu_{ji}}{\tilde{m}_j} \quad \text{and} \quad \Lambda_{j_i^+ i}(t^k) \xrightarrow[k \to \infty]{} \frac{\tilde{m}_{j_i^+ i} \mu_{j_i^+ i}}{\tilde{m}_{j^+}}.$$

We letting J^+ be the set of queues on route i between j and j_i^+ that includes j but does not include j_i^+. Applying a triangle inequality across these queues, we can say that

$$\left| \frac{\tilde{m}_{ji} \mu_{ji}}{\tilde{m}_j} - \frac{\tilde{m}_{j_i^+ i} \mu_{j_i^+ i}}{\tilde{m}_{j^+}} \right| \leq \lim_{k \to \infty} \left[\left| \frac{\tilde{m}_{ji} \mu_{ji}}{\tilde{m}_j} - \Lambda_{ji}(t^k) \right| + \left| \frac{\tilde{m}_{j_i^+ i} \mu_{j_i^+ i}}{\tilde{m}_{j_i^+}} - \Lambda_{j_i^+ i}(t^k) \right| + \sum_{l \in J^+} \left| \Lambda_{li}(t^k) - \Lambda_{x_{li} i}(t^k) \right| \right] = 0.$$

Applying this triangle inequality once more, for any queue l on route i that is between j and j_i^+, we see that

$$\lim_{k \to \infty} \left| \frac{\tilde{m}_{ji} \mu_{ji}}{\tilde{m}_j} - \Lambda_{li}(t^k) \right| = 0.$$

In other words, for each route i and for all queue $j \in i$, $\Lambda_{ji}(t^k)$ converges to some value $\tilde{L}_i > 0$ where if $\tilde{m}_j > 0$ we have that

$$\tilde{L}_i = \frac{\tilde{m}_{ji} \mu_{ji}}{\tilde{m}_j}$$

for some constant $\tilde{L}_i > 0$. Observe that, by (2.10b), for each queue $j \in \mathcal{J}$

$$\sum_{i: j \in i} \frac{\tilde{L}_i}{\mu_{ji}} \leq 1$$

also

$$\beta(\tilde{m}) = \sum_{i \in \mathcal{I}} \sum_{j: j \in i} \tilde{m}_{ji} \log \tilde{L}_i = \sum_{i \in \mathcal{I}} n_i \log \tilde{L}_i. \tag{4.33}$$

Thus, the vector $\tilde{L} = (\tilde{L}_i : i \in \mathcal{I})$ is feasible for the dual problem and the vector $\tilde{m}$ is feasible for the primal problem. We know by Weak Duality (for a minimization) that any primal feasible solution is bigger than that of the dual. Thus, we know by (4.33) that the primal equals dual solution and so $\tilde{m}$ must be optimal for the primal problem i.e. $\beta(\tilde{m}) = \beta(m^*)$. This must be a contradiction because by assumption $\beta(m(t^k)) \geq \beta(m^*) + \epsilon$ and thus by continuity of β, $\beta(\tilde{m}) \geq \beta(m^*) + \epsilon$. □

We are now in a position to prove Theorem 4.

Proof. Proof of Theorem 4 We found $\beta(m(t))$ was absolutely continuous in t. In Lemma 10, we found the derivative of $\beta(m(t))$ was almost everywhere negative and thus $\beta(m(t))$ must be a decreasing function.

Suppose for $s \in [0, t]$, $\beta(m(s)) \geq \beta(m^*) + \epsilon$ for some $\epsilon > 0$ then by Lemma 12 there exists an $i \in \mathcal{I}$ and a $j \in i$

$$|\Lambda_{ji}(s) - \Lambda_{ji}(s)| \geq \delta_\epsilon,$$

for some $\delta_\epsilon > 0$. Thus, applying this to our bound in Lemma 10 for intervals of time $[0, t]$ such that $\beta(m(s)) \geq \beta(m^*) + \epsilon$, we have that

$$\beta(m(t)) \leq \beta(m(0)) - t \frac{\delta_\epsilon^2}{|\mathcal{J}| \mu_{max}}.$$

As $\beta(m(t))$ is bounded below by $\beta(m^*)$, the above inequality cannot be sustained for all times t. In other words, eventually $\beta(m(t)) \leq \beta(m^*) + \epsilon$. Thus $\beta(m(t)) \searrow \beta(m^*)$. This proves the first assertion in Theorem 4.

Now, it remains to show that $m(t)$ approaches $\mathcal{M}$, the set of solutions to (4.32). Take some $\epsilon_1 > 0$. Let $m = (m_{ji} : i \in \mathcal{I}, j \in i)$ be any vector with $\sum_{j: j \in i} m_{ji} = n_i$ for $i \in \mathcal{I}$ and such that

$$\min_{m^* \in \mathcal{M}} |m - m^*| \geq \epsilon_1.$$

Such an m belongs to a compact set and thus, as β is continuous, it must be that $\beta(m) \geq \beta(m^*) + \epsilon$ for some $\epsilon > 0$. Or stated differently, if $\beta(m) < \beta(m^*) + \epsilon$ then it must be that

$$\min_{m^* \in \mathcal{M}} |m - m^*| < \epsilon_1.$$

As we have just shown, $\beta(m(t)) < \beta(m^*) + \epsilon$ holds eventually for all fluid paths. Thus,

$$\lim_{t \to \infty} \min_{m^* \in \mathcal{M}} |m(t) - m^*| = 0.$$

□

Appendix

A Proof of Lemma 2.

We consider both optimal solutions $\bar{\Lambda}^*(n)$ and $\Lambda^*(n)$. Let

$$G_n(\Lambda) = \sum_{i\in\mathcal{I}} n_i \log \Lambda_i$$

Since $\bar{\Lambda}^*(n)$ is the solution of an optimization with a larger feasible set $G_n(\bar{\Lambda}^*(n)) \geq G_n(\Lambda^*(n))$.

Take $v = \bar{\Lambda}^*(n) - \Lambda^*(n)$. Note $\Lambda^*(n) + \delta v$ belongs to feasible set

$$\left\{\Lambda \geq 0 : \sum_{i:j\in i} \frac{\Lambda_i}{\mu_{ji}} \leq 1,\ j \in \mathcal{J}\right\}$$

for all δ suitably small. If this were not so then there would have been some constraint/queue which we did not correctly include in the set of bottleneck links $\bar{J}$. Taking the partial derivative of G_n from $\Lambda^*(n)$ in the direction of v, we can then say that

$$\sum_{i\in\mathcal{I}} v_i \frac{\partial G_n(\Lambda^*(n))}{\partial \Lambda_i} \leq 0.$$

This holds because $\Lambda^*(n)$ is optimal. Now, also, by the concavity of $G_n(\cdot)$

$$G_n(\bar{\Lambda}^*(n)) - G_n(\Lambda^*(n)) \leq \sum_{i\in\mathcal{I}} v_i \frac{\partial G_n(\Lambda^*(n))}{\partial \Lambda_i}.$$

So, $G_n(\bar{\Lambda}^*(n)) \leq G_n(\Lambda^*(n))$, and thus $G_n(\Lambda^*(n)) = G_n(\bar{\Lambda}^*(n))$. By the strict concavity the optimum of $G_n(\cdot)$ is unique, so, it must be that $\bar{\Lambda}^*(n) = \Lambda^*(n)$.

B Lipschitz Continuity of $\beta(m(t))$.

Before proceeding to prove the Lipschitz continuity of $\beta(m(t))$, i.e., Proposition 6, we give a proof of Lemma 7.

Proof. Proof of Lemma 7 We use the fact that we know that the derivative exists. First, it is clear that $\frac{df}{dt} = 0$ because

$$\frac{df}{dt} = \lim_{h\searrow 0} \frac{f(t+h) - 0}{h} \geq 0 \quad \text{and} \quad \frac{df}{dt} = \lim_{h\nearrow 0} \frac{f(t+h) - 0}{h} \leq 0.$$

Now, noting that $f \log(p)$ is negative for all $p \leq 1$ and by using the same argument, we have

$$\frac{df \log p}{dt} = \lim_{h\searrow 0} \frac{f(t+h)\log p(t+h) - 0}{h} \leq 0, \quad \text{and} \quad \frac{df \log p}{dt} = \lim_{h\nearrow 0} \frac{f(t+h)\log p(t+h) - 0}{h} \geq 0.$$

Thus $\frac{df \log f}{dt} = 0$. □

We now demonstrate that the function $\beta(m(t))$ is Lipschitz continuous on any compact time interval. Here $m(t)$ is any solution to the fluid equations (2.10) and $\beta(m)$ is defined by (3.6). The following arguments are adapted from Lemma 4.2 and Proposition 4.2 of Bramson [7]. All queues may empty in our network, so we have to apply some degree of care in proving the Lipschitz continuity on compact time intervals. First, we prove Lemma 8.

Proof. Proof of Lemma 8 Suppose that $m(t)$ and $L(t)$ are differentiable at t. We may assume $\Lambda_{ji}(t) > 0$ for some queue j on route i. Such a queue must exist because there is always some queue with $m_{ji}(t) > 0$ as $n_i > 0$ and thus by (2.10d) $\Lambda_{ji}(t) > 0$. Now consider x_{ji} the next queue on route i, if $m_{x_{ji}i}(t) > 0$ then by (2.10d) $\Lambda_{x_{ji}i} > 0$, and if $m_{x_{ji}i}(t) = 0$ then $m'_{x_{ji}i}(t) = 0$, thus by (2.10a) $\Lambda_{x_{ji}i}(t) = \Lambda_{ji}(t) > 0$. Continuing inductively we see that $\Lambda_{ji}(t) > 0$ for all queues.

From this argument we now see that the value of $\Lambda_{ji}(t)$ on any route i is achieved by a queue j^* with $m_{j^*i} > 0$. Thus applying (2.10d), $\Lambda_{ji}(t) \leq \mu_{ji} \leq \mu_{max}$. □

Proposition 8. *For almost every $t > t_0 > 0$, there exists a constant $T > 0$ such that if $t - t_0 < \kappa T$ for $\kappa \in \mathbb{N}$ then*

$$\min_{ji} \Lambda_{ji}(t) > \frac{\min_{ji} \Lambda_{ji}(t_0)}{(1 + \max_{ji} \mu_{ji})^\kappa}$$

Here κ is a strictly positive constant which depends on $m(t_0)$, the fluid model state at time t_0.

In order to prove this proposition we require the following lemma

Lemma 13. *For almost every time t_0 and t with $t > t_0$, if a queue j on route i, has arrival process from the queue before j, l_{ji}, such that for almost every $s \in [t_0, t]$*

$$\Lambda_{l_{ij}j}(s) > c,$$

then the output of route i from queue j satisfies

$$\Lambda_{ji}(t) \geq c' \tag{B.1}$$

where

$$c' = \min\left\{\Lambda_{ji}(t_0), \frac{c}{1 + \max_{ji} \mu_{ji}}\right\}.$$

Proof. Proof

We note that $L_{ij}(t)$ is a Lipschitz function and thus is almost everywhere differentiable. We assume that t and t' are differentiable points where (B.1) is violated. Observe that if $m_{ji}(t) = 0$ then, by (2.10a), $0 = m'_{ji}(t) = \Lambda_{ji}(t) - \Lambda_{l_{ji}i}(t)$. Thus $\Lambda_{ji}(t) = \Lambda_{l_{ji}i}(t) > c > c'$. So it must be that $m_{ji}(t) > 0$. Consequently by (2.10d), $\Lambda_{ji}(s) = \mu_{ji}m_{ji}(s)/m_j(s)$ must be continuous on an interval around t and there must be an open interval around t for which $\Lambda_{ji}(s) < c'$.

For a $\Lambda_{ji}(s)$ to get small we need the total number of departures to be comparable relative to the arrivals. So, we will next argue the contradiction that $\Lambda_{ji}(t)$ cannot enter an interval of time for which $\Lambda_{ji}(s) < c'$ without the average departure rate $(\Lambda_{ji}(t) - \Lambda_{ji}(s))/(t - s)$ being bigger that c'.

We let $\tilde{t}$ be the last time before t for which $L_{ji}(\tilde{t}) \geq c'$. Note $L_{ji}(t_0) \geq c'$, so $\tilde{t}$ is well defined. We use the shorthand $L_{ji}(\tilde{t}, t) = L_{ji}(t) - L_{ji}(\tilde{t})$ and $\Delta_j(\tilde{t}, t) = m_j(t) - m_j(\tilde{t})$. As $m_{ij}(t) > 0$, by (2.10d), we have

$$c' > \Lambda_{ji}(t) = \mu_{ji}\frac{m_{ji}(t)}{m_j(t)} = \mu_{ji}\frac{m_{ji}(\tilde{t}) + L_{l_{ji}i}(\tilde{t}, t) - L_{ji}(\tilde{t}, t)}{m_j(\tilde{t}) + \Delta_j(\tilde{t}, t)}.$$

Rearranging the above expression implies

$$L_{ji}(\tilde{t}, t) > L_{l_{ji}i}(\tilde{t}, t) - c'\Delta_j(\tilde{t}, t) + \mu_{ji}m_{ji}(\tilde{t}) - c'm_j(\tilde{t}) \geq L_{l_{ji}i}(t) - c'\Delta_j(\tilde{t}, t).$$

The last inequality holds because, by (2.10d), $\Lambda_{ji}(\tilde{t}) \geq c'$ implies $\mu_{ji}m_{ji}(\tilde{t}) - c'm_j(\tilde{t}) \geq 0$. We now look at the mean value of the terms in the above inequality:

$$\frac{L_{ji}(\tilde{t}, t)}{t - \tilde{t}} > \frac{L_{l_{ji}i}(\tilde{t}, t)}{t - \tilde{t}} - c'\frac{\Delta_j(\tilde{t}, t)}{t - \tilde{t}} \geq c - c' \max_{ji} \mu_{ji} \geq c'.$$

In the second inequality above, we use the assumption that $\Lambda_{l_{ij}j}(s) > c,\ s \in [t_0, t_0 + T]$ and the fact the average change in $m_j(t)$, $\Delta_j(\tilde{t}, t)/(t - \tilde{t})$, is at most the maximum service rate $\max_{ji} \mu_{ji}$. The final inequality holds by our choice of c'. This then contradicts our assumption that $\Lambda_{ji}(s) < c'$ on the interval $(\tilde{t}, t]$. ☐

The following lemma is a consequence achieved by iteratively applying the last result. We have to apply some care because, in comparison to Bramson's open network analysis [7], every queue can empty or have low arrival and departure rate.

Lemma 14. *For almost every t_0, there exists an interval of fixed length $[t_0, t_0 + T]$ such that for almost every $t \in [t_0, t_0 + T]$ and for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$*

$$\Lambda_{ji}(t) > \tilde{c}(t_0) > 0.$$

Here $\tilde{c}(t_0)$ is a function of the network state at time t_0.

Proof. Proof On route i there is always one queue with greater than or equal to the average amount of work. Without loss of generality, i.e. relabelling queues if necessary, we assume that this is the first queue on route i. So, we have that for any route i

$$m_{j_i^1 i}(t_0) \geq \frac{\min_r n_r}{I}.$$

The biggest rate that this queue could decrease is $\max_{ji} \mu_{ji}$. So, we have

$$m_{j_i^1 i}(t) \geq \frac{\min_r n_r}{2I} \qquad \text{for} \quad t \in [t_0, t_0 + T],$$

where we define $T = \frac{\min_r n_r}{2I \max_{ji} \mu_{ji}}$, and

$$L_{ji}(t_0) = \mu_{ji}\frac{m_{ji}(t_0)}{m_j(t_0)} > \frac{\min_{ji} \mu_{ji} \min_r n_r}{2I \sum_r n_r} = c_1.$$

We can now repeatedly apply Lemma 13. Starting from c_1 above for $k = 1, \dots, k_i - 1$ we define

$$c_{k+1} = \min\left\{\Lambda_{ji}(t_0), \frac{c_k}{1 + \max_{ji} \mu_{ji}}\right\}.$$

Each time we iterate, we reduce c_k by at least $1+\max_{ji}\mu_{ji}$. A simple lower bound, which is sufficient for our purposes, is that

$$c_k \geq \frac{\min_{ji}\Lambda_{ji}(t_0)}{(1+\max_{ji}\mu_{ji})^K} \stackrel{\text{def}}{=} \tilde{c}(t_0)$$

where $K=\max_i k_i$ is the longest route within the queueing network. Thus, from this repeated application of Lemma 13, we have that

$$\Lambda_{ji}(t) \geq \tilde{c}(t_0)$$

for almost every $t \in [t_0, t_0+T]$. □

In a similar manner to Proposition 4.2 of Bramson [7], now we show that $\beta(m(t))$ is Lipschitz on any compact time interval.

Proof. Proof of Proposition 6 Note that without loss of generality, we may assume interval $[t_0, t]$ is of length less than or equal to $T = \frac{\min_r n_r}{2I\max_{ji}\mu_{ji}}$, where T was derived in the last lemma, Lemma 14. If $t-t_0>T$ then we can split the interval $[t_0, t]$ in to overlapping sub-interval of size T and then use the largest Lipschitz constant found in each sub-interval as a Lipschitz constant for $[t_0.t]$.

Note that

$$\beta(m(t)) = \sum_{j\in\mathcal{J}} \sum_{\substack{i:j\in i\\ m_{ji}(t)>0}} m_{ji}(t)\log\frac{m_{ji}(t)\mu_{ji}(t)}{m_j(t)} = \sum_{j\in\mathcal{J}}\sum_{i:j\in i} m_{ji}(t)\log\Lambda_{ji}(t).$$

It is enough to prove Lipschitz continuity of each term summed above:

$$|m_{ij}(t_2)\log(\Lambda_{ji}(t_2)) - m_{ij}(t_1)\log(\Lambda_{ji}(t_1))| \leq D_1\,|t_2-t_1|\ .$$

By Lemma 14 for $s\in[t_0,t]$, $\log(\Lambda_{ji}(s))$ is bounded below by $\tilde{c}(t_0)$ and also above by m_j, the maximum service rate of queue j. So

$$|\log(\Lambda_{ji}(s))| \leq D_0 \stackrel{\text{def}}{=} \max\{|\log(m_j)|, |\log(\tilde{c}(t_0))|\}.$$

If $m_{ij}(t_2)=m_{ij}(t_1)=0$, the relation is trivial, so we assume that $m_{ij}(t_2)>0$. If $m_{ij}(t_1)=0$ we have that

$$|m_{ij}(t_2)\log(\Lambda_{ji}(t_2))| \leq D_0|m_{ij}(t_2)| = D_0|m_{ij}(t_2)-m_{ij}(t_1)| \leq D_2\,D_0\,|t_2-t_1|$$

where the constant D_0 is as above and the constant D_2 is the Lipschitz constant of $m_{ij}(t)$.

Now assume $m_{ij}(t_2)$ and $m_{ij}(t_1)$ both positive, and without loss of generality that

$$\Lambda_{ji}(t_1) \leq \Lambda_{ji}(t_2)\ .$$

It follows that

$$\begin{aligned}|m_{ij}(t_2)\log\Lambda_{ji}(t_2) - m_{ij}(t_1)\log\Lambda_{ji}(t_1)| \leq{}& |m_{ij}(t_2)-m_{ij}(t_1)|\ |\log\Lambda_{ji}(t_2)| \\ &+ m_{ij}(t_1)\ |\log\Lambda_{ji}(t_2) - \log\Lambda_{ji}(t_1)|\end{aligned}$$

Again the first term in this upper bound is less than or equal to $D_2\,D_0\,|t_2-t_1|$. The second term, since the logarithm function is concave and has its derivative maximized at the left-most point, can be bounded in the following way

$$\begin{aligned}
m_{ij}(t_1)\ |\log\Lambda_{ji}(t_2) - \log\Lambda_{ji}(t_1)| &\leq \frac{m_{ij}(t_1)}{\Lambda_{ji}(t_1)}\,|\Lambda_{ji}(t_2)-\Lambda_{ji}(t_1)| \\
&= \left| m_j(t_1)\frac{m_{ji}(t_2)}{m_j(t_2)} - m_{ij}(t_1)\right| \\
&\leq \frac{m_{ij}(t_2)}{m_j(t_2)}\,|m_j(t_2)-m_j(t_1)| + |m_{ij}(t_2)-m_{ij}(t_1)| \\
&\leq |m_j(t_2)-m_j(t_1)| + |m_{ij}(t_2)-m_{ij}(t_1)| \\
&\leq D_3\,|t_2-t_1|,
\end{aligned}$$

which follows from the fact that $m_{ij}(t_2) \leq m_j(t_2)$ and the fact that both $m_{ij}(t)$ and $m_j(t)$ are Lipschitz continuous. The result follows by choosing $D_1 \geq \max\{D_3, D_2\,D_0\}$. □

References

[1] J. Anselmi and P. Cremonesi, *A unified framework for the bottleneck analysis of multiclass queueing networks*, Perform. Eval. **67** (2010), 218–234.

[2] J. Anselmi, P. Cremonesi, and E. Amaldi, *Service consolidation with end-to-end response time constraints*, Euromicro-SEAA, IEEE, 2008, pp. 345–352.

[3] G. Balbo and G. Serazzi, *Asymptotic analysis of multiclass closed queueing networks: Multiple bottlenecks*, Performance Evaluation **30** (1997), no. 3, 115–152.

[4] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios, *Open, closed, and mixed networks of queues with different classes of customers.*, Journal of the ACM **22** (1975), no. 2, 248–260.

[5] A. Berger, L. Bregman, and Y. Kogan, *Bottleneck analysis in multiclass closed queueing networks and its application*, Queueing Syst. Theory Appl. **31** (1999), no. 3-4, 217–237.

[6] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi, *Queueing networks and markov chains*, Wiley-Interscience, 2005.

[7] Maury Bramson, *Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks*, Queueing Systems **23** (1996), 1–26.

[8] T. C. Brown and P. K. Pollett, *Some distributional approximations in markovian queueing networks*, Advances in Applied Probability **14** (1982), no. 3, pp. 654–671.

[9] G. Casale, *A generalized method of moments for closed queueing networks*, Perform. Eval. **68** (2011), no. 2, 180–200.

[10] G. Casale and G. Serazzi, *Bottlenecks identification in multiclass queueing networks using convex polytopes*, MASCOTS '04 (Washington, DC, USA), IEEE Computer Society, 2004, pp. 223–230.

[11] K. Mani Chandy and Doug Neuse, *Linearizer: a heuristic algorithm for queueing network models of computing systems*, Commun. ACM **25** (1982), no. 2, 126–134.

[12] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks*, Springer-Verlag, 2001.

[13] Hong Chen and Avi Mandelbaum, *Stochastic discrete flow networks: diffusion approximations and bottlenecks*, Ann. Probab. **19** (1991), no. 4, 1463–1519.

[14] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley-Interscience, August 1991.

[15] P. Dupuis and M. Fischer, *On the construction of lyapunov functions for nonlinear markov processes via relative entropy*, preprint (2012).

[16] David K. George, Cathy H. Xia, and Mark S. Squillante, *Exact-order asymptotic analysis of closed queueing networks*, Journal of Applied Probability (2012).

[17] Jonathan B. Goodman and William A. Massey, *The non-ergodic jackson network*, Journal of Applied Probability **21** (1984), no. 4, pp. 860–869.

[18] W. J. Gordon and G.F. Newell, *Closed queuemg systems with exponential servers*, Operation Research (1967), no. 15, 254–265.

[19] Peter G. Harrison and Sergio Coury, *On the asymptotic behaviour of closed multiclass queueing networks*, Perform. Eval. **47** (2002), no. 2, 131–138.

[20] Haya Kaspi and Avi Mandelbaum, *Regenerative closed queueing networks*, Stochastics Stochastics Rep. **39** (1992), no. 4, 239–258.

[21] F. P. Kelly, *Charging and rate control for elastic traffic*, European Transactions on Telecommunications **8** (1997), 33–37.

[22] F. P. Kelly, L. Massoulié, and N. S. Walton, *Resource pooling in congested networks: proportional fairness and product form*, Queueing Syst. Theory Appl. **63** (2009), 165–194.

[23] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, *Rate control in communication networks: shadow prices, proportional fairness and stability*, Journal of the Operational Research Society **49** (1998), 237–252.

[24] F.P. Kelly, *Reversibility and stochastic networks*, Wiley, Chicester, 1979.

[25] Charles Knessl and Charles Tier, *Asymptotic expansion for large closed queuing networks*, Journal of the ACM **37** (1990), no. 1, 144–174.

[26] ———, *Asymptotic expansions for large closed queueing networks with multiple job classes*, IEEE Trans. Comput. **41** (1992), no. 4, 480–488.

[27] P. R. Kumar and T. I. Seidman, *Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems*, IEEE Transactions on Automatic Control **35** (1990), 289–298.

[28] Lester Lipsky, Chee-Min Henry Lieu, Abolfazl Tehranipour, and Appie van de Liefvoort, *On the asymptotic behavior of time-sharing systems*, Commun. ACM **25** (1982), no. 10, 707–714.

[29] J. McKenna and D. Mitra, *Asymptotic expansions and integral representations of moments of queue lengths in closed markovian networks*, Journal of the ACM **31** (1984), no. 2, 346–360.

[30] J. McKenna and Debasis Mitra, *Asymptotic expansions for closed markovian networks with state-dependent service rates*, J. ACM **33** (1986), no. 3, 568–592.

[31] K. R. Pattipati, M. M. Kostreva, and J. L. Teele, *Approximate mean value analysis algorithms for queuing networks: existence, uniqueness, and convergence results*, J. ACM **37** (1990), no. 3, 643–673.

[32] B. Pittel, *Closed exponential networks of queues with saturation: The jackson-type stationary distribution and its asymptotic analysis*, Mathematics of Operations Research **4** (1979), no. 4, 357–378.

[33] P.K Pollett, *Modelling congestion in closed queueing networks*, International Transactions in Operational Research **7** (2000), no. 45, 319 – 330.

[34] M. Reiser and H. Kobayashi, *Queueing networks with multiple closed chains: Theory and computational algorithms*, IBM J. Res. Dev. **19** (1975), no. 3, 283–294.

[35] M. Reiser and S. S. Lavenberg, *Mean-value analysis of closed multichain queueing networks*, Journal of the ACM **27** (1980), no. 2, 313–322.

[36] Philippe Robert, *Stochastic networks and queues*, Springer-Verlag, 2000.

[37] P. Schweitzer, *Approximate analysis of multi-class closed networks of queues*, In Proceedings of the International Conference on Stochastic Control and Optimization (1979).

[38] ______, *Bottleneck determination in networks of queues*, Proc. ORSA/TIMS Special Interest Conf on Appl. Probab. - Comput. Sci., The Interface, Boca Raton, FLA (1981), 471–485.

[39] P. Schweitzer, G. Serazzi, and M. Broglia, *A survey of bottleneck analysis in closed queues*, Perf. Eval. of Comp. and Comm. Sys., LNCS, No. 729, Springer-Verlag, Berlin (1993), 491–508.

[40] F. Spitzer and Mathematical Association of America, *Random fields and interacting particle systems: notes on lectures given at the 1971 maa summer seminar, williams college, williamstown, massachusetts*, Random Fields and Interacting Particle Systems, Mathematical Association of America, 1971.

[41] R. Srikant, *The mathematics of internet congestion control*, Birkhauser, 2004.

[42] Bhuvan Urgaonkar, Giovanni Pacifici, Prashant Shenoy, Mike Spreitzer, and Asser Tantawi, *An analytical model for multi-tier internet services and its applications*, ACM SIGMETRICS, ACM, 2005, pp. 291–302.

[43] N. S. Walton, *Proportional fairness and its relationship with multi-class queueing networks*, Annals of Applied Probability **19** (2009), no. 6, 2301–2333.

[44] H. Wang, K. Sevcik, G. Serazzi, and S. Wang, *The general form linearizer algorithms: A new family of approximate mean value analysis algorithms*, Perform. Eval. **65** (2008), no. 2, 129–151.

[45] W. Whitt, *Open and closed models for networks of queues*, AT and T Bell Laboraties Technical Journal **63** (1984), no. 9, 1911–1979.