

Validity of heavy-traffic steady-state approximations in multiclass queueing networks: The case of queue-ratio disciplines

Itai Gurvich

Kellogg School of Management, Northwestern University, Evanston, IL 60208
email: i-gurvich@kellogg.northwestern.edu

A class of stochastic processes known as semi-martingale reflecting Brownian motions (SRBMs) is often used to approximate the dynamics of heavily loaded queueing networks. In two influential papers, Bramson (1998) and Williams (1998) laid out a general and structured approach for proving the validity of such heavy-traffic approximations, in which an SRBM is obtained as a diffusion limit from a sequence of suitably normalized workload processes. However, for multiclass networks it is still not known in general whether the *steady-state distribution* of the SRBM provides a valid approximation for the steady-state distribution of the original network. In this paper we study the case of queue-ratio disciplines and provide a set of sufficient conditions under which the above question can be answered in the affirmative. In addition to standard assumptions made in the literature towards the stability of the pre- and post-limit processes and the existence of diffusion limits, we add a requirement that solutions to the fluid model are attracted to the invariant manifold at linear rate. For the special case of static-priority networks such linear attraction is known to hold under certain conditions on the network primitives. The analysis elucidates interesting connections between stability of the pre- and post-limit processes, their respective fluid models and state-space collapse, and identifies the respective roles played by all of the above in establishing validity of heavy-traffic steady-state approximations.

Key words: Steady-state ; Multi-class ; Heavy traffic ; Network ; Queue-ratio

MSC2000 Subject Classification: Primary: 60K25, 90B15 ; Secondary: 60F17 , 60J20

OR/MS subject classification: Primary: Queueing networks , Limit theorems ; Secondary: Markov processes , Diffusion models

1. Introduction and overview of the main contribution

1.1 Motivation and the main question Queueing networks are commonly used to model communication networks and complex service and manufacturing systems. In many cases more than a single class of jobs can be processed at each station, and the model is then collectively referred to as a *multiclass* queueing network. These models represent a significant escalation in complexity relative to their single class counterparts and, for all but the simplest cases, are rarely amenable to exact analysis.

In an effort to establish tractable representations for these types of complex systems, much of the research on stochastic processing networks has focused on approximate analysis. The most prevalent types of approximations found in the literature fall into the following two categories: (i) *fluid approximations* that are mostly used for stability analysis; and (ii) *diffusion approximations* that are used for performance analysis of heavily-loaded systems.

Deriving diffusion approximations for queueing networks has been the focus of research since the early 60's; see, e.g. [27, 24, 22, 31]. The standard formulation considers a sequence of systems in which time and space are scaled in accordance with the functional central limit theorem, and the traffic intensity (utilization) is made to approach 1 at a suitable rate (for this reason these are often referred to as heavy-traffic approximations). The seminal papers by Bramson [5] and Williams [38] provide a broad set of sufficient conditions for the validity of such diffusion approximations for multiclass queueing networks. In particular, Williams [38] proves that as the traffic intensity approaches one, the normalized vector of queue length processes converges to a diffusion process known as a semi-martingale reflecting Brownian motion (SRBM). This SRBM is often referred to as the “Brownian model” or “Brownian counterpart” of the original queueing network.

The main appeal of the Brownian system model is that it provides a relatively tractable and rigorous approximation for the queue length *dynamics*. In addition, one can use the stationary distribution of the SRBM as a scaled proxy for the *steady-state* behavior of the underlying queueing network. The advantages of this approach are evident: the steady-state behavior of the original queueing network can typically only be characterized via exhaustive simulation, while the SRBM is a diffusion process whose stationary distribution can be obtained by solving partial differential equations. While these equations

will typically not give rise to closed-form expressions for the stationary distribution, they can nonetheless be solved relatively efficiently using a variety of numerical algorithms; see, e.g., Dai and Harrison [15], Chen and Shen [9], and Saure, Glynn and Zeevi [33].

The use of a Brownian system model as a means to approximate the network’s steady-state distribution has been advocated by several papers in the literature. Harrison and Nguyen [23] formalized this procedure, articulating an approximation scheme named QNET. The first step in QNET constructs the Brownian system model from the problem primitives characterizing the original network. Then, the steady-state workload in the queueing network is approximated by that of the Brownian model (suitably scaled). While this approximation is clearly motivated by heavy-traffic theory, there is no rigorous justification for the transition from approximations over finite time intervals (the diffusion limits) to an approximation over an infinite time horizon (steady-state variables).

To better explain the main issues underlying validity of heavy-traffic steady-state approximations, consider a sequence of queueing networks indexed by r that satisfy the following *heavy-traffic condition*:

$$\sqrt{r}(1 - \rho_j^r) \rightarrow \gamma_j \quad \text{as } r \rightarrow \infty, \quad (1)$$

for each station j , where ρ_j^r is the utilization in station j and γ_j is a positive constant; a more precise definition will be given in §2.1. Let $\widehat{Z}^r(t) = Z^r(rt)/\sqrt{r}$ denote the properly scaled queue-length vector in the r^{th} network at time $t \geq 0$. To justify a Brownian approximation of the steady-state distribution one must prove the following *limit-interchange*:

$$\lim_{r \rightarrow \infty} \lim_{t \rightarrow \infty} \mathbb{E} \left[f(\widehat{Z}^r(t)) \right] = \lim_{t \rightarrow \infty} \lim_{r \rightarrow \infty} \mathbb{E} \left[f(\widehat{Z}^r(t)) \right], \quad (2)$$

for all bounded and continuous functions f . This is expressed graphically in Figure 1.

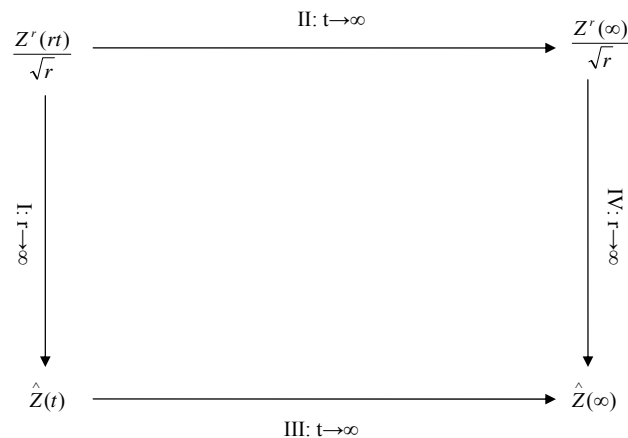


Figure 1: The interchange-of-limits diagram

The figure has four components: (I) diffusion limits (process convergence); (II) stability and existence of a steady-state for the pre-limit; (III) stability and existence of a steady-state for the Brownian model; and (IV) convergence of the steady-state distributions.

To date, this limit-interchange problem in open queueing networks has only been worked out for networks with a single customer class otherwise known as generalized Jackson networks. In particular, the recent papers by Gamarnik and Zeevi [20] and, subsequently, Budhiraja and Lee [6] derived such a result, and consequently established the validity of the Brownian steady-state approximation for this class of networks. It is worth pointing out that in the context of that problem, edges (I), (II) and (III) were known, and the work in [20] and [6] established (IV), hence proving that the limit interchange (2) is valid.

The limit interchange has been established for some instances of multiclass queueing systems in heavy-traffic. Katsuda [26] (further discussed towards the end of this section) proves the limit interchange result for both the queue-length and workload processes in a multiclass single-server queue with feedback under

various disciplines. Ye and Yao [40] study a parallel-server system with two customer classes and two servers. Gamarnik and Stolyar [19] and Tezcan [35] prove the limit-interchange for special instances of parallel-server systems in the so-called Halfin-Whitt heavy-traffic regime.

In this paper we add to the existing results in studying multiclass open queueing networks with queue-ratio disciplines. We will refer to these as *queue-ratio networks*. Queue-ratio disciplines seek to set the queues at each station equal to a linear combination of the queue in that station. These ratios can be arbitrarily set, rendering this a fairly general family of disciplines. The policies are explicitly defined in §2.2.

1.2 Connections to antecedent literature and summary of the paper’s main contribution

Queue-ratio networks present new and significant challenges that did not appear in the case of generalized Jackson networks. While we are not able to develop a complete theory that encompasses all disciplines for which process-level convergence to an SRBM limit (edge (I) in Figure 1) has been rigorously verified, we develop a systematic approach and, for the family of queue-ratio disciplines, identify a simple set of sufficient conditions.

To be a bit more specific, yet speaking very loosely at this stage, our main result states that given a sequence of stable queue-ratio networks, whose properly normalized queue-length vector converges to an SRBM that is itself stable, the limit interchange is valid if solutions to the fluid model are attracted to the so-called invariant manifold *at a linear rate*.

All of this will be carefully explained in what follows, but we will note that the theoretical constructs we use and develop build heavily on, and present interesting connections to, the key papers in the field that have established the existing three edges (I-III) of the diagram in Figure 1. We next summarize some of the key ideas related to these edges and the manner in which the current paper builds on that theory.

Fluid models and stability analysis: Determining whether a queueing network is stable, i.e., whether it admits a stationary distribution, is greatly simplified by reducing the problem to the study of stability of a deterministic counterpart known as a *fluid model*. An important result due to Dai [14] (see also Stolyar [34]) shows that if the fluid-model “queues” are emptied in a finite time, then the original queueing network is stable. Fluid models play an analogous role in studying stability of the Brownian counterpart to the original network. Dupuis and Williams [18] show that an SRBM is stable if its fluid model (a deterministic Skorohod Problem) drains to the origin in finite time. Our work builds on the results of Dai [14] and Dupuis and Williams [18]. The latter will play a key role in our analysis, which hinges on identifying a suitable Lyapunov function for the queue-length process. This illustrates an important connection between stability of the SRBM, as viewed through the lens of a fluid model, and establishing tightness of the sequence of steady-state pre-limit queue-length processes.

Diffusion limits and state-space collapse: Up until the work of Williams [38], the standard approach for establishing diffusion limits, within the heavy-traffic framework described earlier, relied on the Continuous Mapping Theorem. This, in turn, hinges on the continuity of an underlying Skorohod mapping; one of the first illustrations of this approach is Reiman’s seminal paper on generalized Jackson networks [31]. The continuous mapping approach was used also in a handful of specific multiclass queueing network settings, such as static-priority feed-forward networks [30], or re-entrant static-priority lines with deterministic routing [11]. However, for all but a very small family of networks, the continuous mapping approach cannot be applied in multiclass settings due to the absence of a path-to-path mapping in the associated Skorohod problem; see [2, 29]. The two main assumptions in Williams [38] are: (a) the regulator matrix R is completely- \mathcal{S} (see §3); and (b) the sequence of networks in heavy-traffic admits a so-called state-space collapse (SSC) property. The first property is necessary for the existence of an SRBM process. The second property guarantees that the queue-length vector (whose dimension is equal to the number of job classes) is given, in the limit, as a linear mapping of the workload process (whose dimension is given by the number of stations). In other words, it is assumed that there exists a matrix Δ , so that, uniformly on compact sets, as $r \rightarrow \infty$,

$$\widehat{Z}^r - \Delta \widehat{W}^r \Rightarrow 0, \tag{3}$$

where \widehat{Z}^r and \widehat{W}^r are the properly scaled queue-length and workload vectors. Consequently, in the limit the state-space *collapses* into one of lower dimension and the matrix Δ is therefore referred to as a *lifting*

matrix. The limit process is then said to live on an *invariant manifold*. The queue-ratio disciplines that we study here use only queue length information (rather than workload), yet a version of (3) remains central to the analysis after replacing the scaled workload \widehat{W}^r with an appropriate linear combination of the queues; see §2.2.

SSC has been established for specific cases (see, for example, Whitt [36], Reiman [32]), but a unified framework was first provided by Bramson [5]. There, conditions for SSC are spelled out in terms of attraction of the fluid model to the so called invariant manifold. The SSC assumption is central to the proofs of Williams [38] and, together with certain Oscillation inequalities, fills gaps created by the absence of a continuous mapping.

Static-priority networks are a special case of queue-ratio networks. Diffusion limits for static priority networks (building on state space collapse and the framework in [38]) have been established in a sequence of papers by Chen and co-authors [10, 11, 12] where explicit conditions are also provided for *linear* attraction of the fluid model to the invariant manifold. We impose such a linear attraction as a condition towards limit interchange.

SSC also plays an instrumental role in our approach to the limit-interchange problem. While we adopt a more stringent notion of state-space collapse, we build heavily on Bramson’s framework and in particular on the connections between SSC and the network’s fluid model. One of the key steps in proving validity of heavy traffic steady-state approximations is to show that SSC holds for suitable sequences of steady-state quantities. For this we introduce a *truncated* analogue of the fluid model. The truncated fluid model allows to prove SSC in steady-state before (and independently of) proving the tightness of the scaled steady-state queues.

Convergence of the steady-state distributions: Provided that a diffusion limit is proved (I), and that the stability of the queueing network (II) and the Brownian model (III) have been established, it suffices to show that the sequence $\{\widehat{Z}^r(\infty)\}$ is tight in order to prove the limit-interchange result for queue-ratio networks. As indicated earlier, this has been established in the case of the single class generalized Jackson networks in [20] and [6]. While the two papers differ somewhat in terms of methodology, both rely on the continuous mapping approach which can not be directly extended to the multiclass case that we consider in the current paper. Our analysis does, however, draw on [20], at least in terms of Lyapunov function arguments. It is also worth pointing out that recent work of Katsuda [26] has shown for a large family of multiclass queueing networks that, *provided* that the sequence of scaled steady-state queues $\{\widehat{Z}^r(\infty)\}$ or the sequence of steady-state workloads $\{\widehat{W}^r(\infty)\}$ are tight, the results of [5] and [38] can be extended to the case in which one initializes the system at time $t = 0$ with its steady-state distribution. In terms of disciplines, our scope is more limited. The main focus of our paper is on proving that, for networks operated under queue-ratio disciplines, the sequence of steady-state queues $\{\widehat{Z}^r(\infty)\}$ is indeed tight provided that a linear attraction condition holds for related fluid models.

Summary of the paper’s contributions: The sufficient conditions in our main result, which is given in §3, reduce the question of limit-interchange in multiclass queue-ratio networks to properties of fluid-models. Recall that if the fluid model corresponding to the pre-limit network is stable in the sense of Dai [14], and if the conditions in Williams [38] hold (namely, SSC in the sense of Bramson [5] and the regularity of the reflection matrix), and the “fluid model” of the corresponding SRBM is stable in the sense of Dupuis and Williams [18], then one has edges (I-III) of the interchange diagram. We add to this by identifying a condition that guarantees that the interchange (IV) holds. The main technical steps that are used to establish this claim boil down to identifying a suitable Lyapunov function for the Brownian model, and using this Lyapunov function as a *constrained* Lyapunov function for the sequence of queueing networks in heavy-traffic. A steady-state version of state-space collapse (via *truncated* fluid models) and crude preliminary bounds on the steady-state queue length are then combined with the constrained Lyapunov function to show the tightness of the sequence of diffusion-scale steady-state queue lengths.

2. Essential preliminaries

2.1 The network model In this subsection we describe the essential elements of the network model. Our description follows mostly that of Williams [38]. The setting that we consider is more restricted and

we will point out wherever our construction departs from hers.

We consider a queueing network with a set $\mathcal{J} = \{1, \dots, J\}$ of single-server stations and, a set $\mathcal{K} = \{1, \dots, K\}$ of customer classes (with $K \geq J$). The many-to-one mapping from customer classes to stations is described by a $J \times K$ constituency matrix C where for $j \in \mathcal{J}$ and $k \in \mathcal{K}$, $C_{jk} = 1$ if class k is served at station j , and it equals 0 otherwise. For $k \in \mathcal{K}$, we let $s(k)$ be the station at which class k is served, i.e., $s(k)$ is the unique $j \in \mathcal{J}$ such that $C_{jk} = 1$.

For each class $k \in \mathcal{K}$, $E_k = (E_k(t), t \geq 0)$ counts the number of arrivals to class k from outside the network that have occurred by time t . Not all classes have exogenous arrivals but we assume that the set $\mathcal{K}^a = \{k \in \mathcal{K} : E_k \not\equiv 0\}$ is non-empty. For each $k \in \mathcal{K}^a$, E_k is a (possibly delayed) renewal process constructed from a sequence of nonnegative random variables $\{u_k(i), i = 1, 2, \dots\}$, where $u_k(i)$ denotes the time between the $(i-1)^{st}$ and the i^{th} external arrival of a class- k customer so that $u_k(1)$ is the time measured from zero until the first external arrival to class k . It is assumed that $\{u_k(i), i = 2, 3, \dots\}$ is a sequence of positive independent and identically distributed (i.i.d) random variables with distribution $F_k^a(\cdot)$, mean $1/\alpha_k \in (0, \infty)$ and coefficient of variation $c_{a,k} \in [0, \infty)$. (The first residual interarrival time, $u_k(1)$, is allowed to have a different distribution.) To be able to apply the stability results of [14] directly, we further require that the inter-arrival times are unbounded and spreadout (see §1 of [14]).

Letting $U_k(0) = 0$ and $U_k(n) = \sum_{i=1}^n u_k(i)$, for $n = 1, 2, \dots$, the renewal process E_k satisfies, for all $t \geq 0$,

$$E_k(t) = \sup\{n \geq 0 : U_k(n) \leq t\}.$$

For convenience, we define $E_k \equiv 0$ for $k \notin \mathcal{K}^a$ and set $E = \{E_k, k \in \mathcal{K}\}$. In our analysis, we will sometimes initialize the queueing network with its steady-state distribution, in which case $u_k(1)$ will have the equilibrium distribution of the corresponding renewal process.

For each $k \in \mathcal{K}$ we denote by $\{v_k(i), i = 2, 3, \dots\}$ the service-time requirements of jobs in class k in order of their entrance to service, so that $v_k(2)$ is the service time of the first class k customer to commence service after time 0. The random variable $v_k(1)$ stands for the residual service time of the customer at the head of the class- k queue at time 0 if the service of that customer has already begun. We set $v_k(1) = 0$ if there is no such customer. Under preemptive disciplines there may be a customer whose service has begun but is not in service.

It is assumed that $\{v_k(i), i = 2, 3, \dots\}$ is a sequence of positive i.i.d. random variables with distribution $F_k^s(\cdot)$, mean $m_k \in (0, \infty)$ and coefficient of variation $c_{s,k} \in [0, \infty)$. We let M denote the $K \times K$ diagonal matrix with m_k as the k^{th} diagonal element. The parameter $\mu_k = 1/m_k$ then stands for the long-run average rate at which class- k customers would be served if the server in station $s(k)$ were never idle and worked exclusively on class k .

The cumulative-service-time process for class k is defined by $V_k(0) = 0$ and $V_k(n) = \sum_{i=1}^n v_k(i)$, for $n = 1, 2, \dots$, and we define the (possibly delayed) renewal process

$$S_k(t) = \begin{cases} \sup\{n \geq 0 : V_k(n) \leq t\}, & \text{if } v_k(1) > 0, \\ \sup\{n \geq 0 : V_k(n) \leq t\} - 1, & \text{if } v_k(1) = 0. \end{cases}$$

The residual service time of the class- k customer in service at time 0, $v_k(1)$, may have a different distribution. Departing from [38], we assume that the service time of a job is generated *when the server commences processing that job* (as opposed to assuming it is generated upon arrival to the processing station).

For both the interarrival and service times it is assumed that, for all $p \in \mathbb{N}$,

$$\begin{aligned} \sup_{z \in \mathbb{R}_+} \mathbb{E}[(u_k(2) - z)^p | u_k(2) > z] &< \infty, \text{ for all } k \in \mathcal{K}^a, \\ \sup_{z \in \mathbb{R}_+} \mathbb{E}[(v_k(2) - z)^p | v_k(2) > z] &< \infty, \text{ for all } k \in \mathcal{K}. \end{aligned} \tag{4}$$

The routing in the network is assumed to be Markovian with a routing matrix P so that P_{kl} is the probability that a class- k customer becomes a class- l customer upon its completion of service at station $s(k)$. The matrix $\tilde{P} = P'$ denotes the transpose of P . To ensure that our queueing network is open, the matrix P (and, in turn, \tilde{P}) is assumed to have spectral radius less than 1.

More formally, let e_1, \dots, e_K be the unit basis vectors parallel to the K coordinate axes in \mathbb{R}^K , and let e_0 be the K -dimensional vector of all zeros. For each class $k \in \mathcal{K}$, $\{\phi^k(i), i = 1, 2, \dots\}$ is a sequence of i.i.d routing vectors where $\phi^k(i)$ takes values in the set $\{e_0, e_1, \dots, e_K\}$. The i^{th} class- k customer to depart from station $s(k)$ is routed to class l if $\phi^k(i) = e_l$, or it leaves the network if $\phi^k(i) = e_0$. Accordingly, $P_{kl} = \mathbb{P}\{\phi^k(i) = e_l\}$, for $k \in \mathcal{K}, l \in \mathcal{K}$. Then, for $k \in \mathcal{K}$,

$$\mathbb{E}[\phi^k(i)] = \tilde{P}^k \text{ and } \text{Cov}[\phi^k(i)] = \Upsilon^k,$$

where \tilde{P}^k denoted the k^{th} column of \tilde{P} , and Υ^k is the $K \times K$ matrix defined by

$$\Upsilon_{lm}^k = \begin{cases} P_{kl}(1 - P_{kl}) & \text{if } l = m, \\ -P_{kl}P_{km} & \text{if } l \neq m. \end{cases} \quad (5)$$

For each $k \in \mathcal{K}$, we define the K -dimensional cumulative routing process for class k by

$$\varphi^k(n) = \sum_{i=1}^n \phi^k(i), \quad n = 1, 2, \dots,$$

where $\varphi^k(0) = 0$. Since \tilde{P} has spectral radius that is strictly smaller than 1, the matrix

$$Q = (I - \tilde{P})^{-1} = I + \tilde{P} + (\tilde{P})^2 + (\tilde{P})^3 + \dots,$$

where $(\tilde{P})^n$ denotes the n^{th} power of \tilde{P} , is well defined.

Finally, we assume that $\{u_l(i), i = 2, 3, \dots\}$, $\{v_k(i), i = 2, 3, \dots\}$, and $\{\phi^k(i), i = 1, 2, \dots\}$, for $l \in \mathcal{K}^a$ and $k \in \mathcal{K}$ are mutually independent sequences of random variables (or vectors), and that collectively these are independent of $(Z_k(0), v_k(1), u_k(1); k \in \mathcal{K})$, where $Z_k(0)$ is the number of class- k customers present in station $s(k)$ at time $t = 0$. We shall refer to the stochastic processes E, V and φ as the *primitives* for the multiclass queueing network model. We assume that all the random variables and stochastic processes introduced thus far are built on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

2.2 Queue-ratio disciplines and a Markovian state descriptor A *lifting matrix* is a $K \times J$ matrix

$$\Delta_{kj} = \begin{cases} \delta_k & \text{if } s(k) = j, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

where $\delta_k, k \in \mathcal{K}$ are non-negative constants such that $CM\Delta = I$ where I is the identity matrix. Recall that $Z_k(t)$ is the queue length of class k at time t . Given a lifting matrix Δ , we define

$$\epsilon(t) = Z(t) - \Delta CMZ(t).$$

The queue-ratio discipline corresponding to Δ then is defined as follows: at each time t , the customer in service in station j is the customer at the head of the highest index class that has a positive value of $\epsilon(t)$. Namely, it is a customer from class k^* where

$$k^* = \max\{i : C_{ji} = 1, \epsilon_i(t) > 0\}. \quad (7)$$

If there are no such classes, the customer at the head of the highest index non-empty queue at that station is served, i.e.,

$$k^* = \max\{i : C_{ji} = 1, Z_i(t) > 0\}. \quad (8)$$

The transition between jobs is made in a preemptive resume manner. Intuitively, the discipline seeks to render the queue of class k proportional to a linear combination of the queues in station $s(k)$. This motivates the name *queue-ratio* discipline. A queue-ratio discipline can be defined for any lifting matrix Δ . Once specified, this matrix completely defines the discipline.

Preemptive resume static priorities are a special instance of queue-ratio disciplines where, for each station j , $\delta_k = 1/m_k$ if $k = \ell(j) := \min\{i : C_{ji} = 1\}$ and $\delta_k = 0$ otherwise. In turn, $\epsilon_k(t) = Z_k(t)$ for all $k \neq \ell(j)$ and server j serves a class- k customer only if all higher priority queues are empty.

For $k \in \mathcal{K}^a$. let $\mathcal{R}_k^a(t)$ be the residual time until the first class- k exogenous arrival after time t . Put $\mathcal{R}^a = (\mathcal{R}_k^a, k \in \mathcal{K}^a)$. If the service of the customer at the head of the class- k queue at time t has already begun, we denote by $\mathcal{R}_k^v(t)$ its residual service time. If the processing of the head-of-the-line class- k customer has not begun at time t we set $\mathcal{R}_k^v(t) = 0$. Putting $\mathcal{R}^v = (\mathcal{R}_k^v, k \in \mathcal{K})$, let

$$\Xi = (Z, \mathcal{R}^a, \mathcal{R}^v), \quad (9)$$

and let $\mathcal{X} \in \mathbb{N}^K \times \mathbb{R}_+^{|\mathcal{K}^a|} \times \mathbb{R}_+^K$ be the domain on which the process Ξ takes its values.

We let $T = (T(t), t \geq 0)$ be the allocation process so that the k^{th} component of $T(t)$ is the cumulative service time allocated to class k up to time t . Letting $\{\sigma_\ell\}_{\ell=0}^\infty$ be a strictly increasing sequence of times at which successive arrivals or departures occur to or from any class in the network, the process $\dot{T}(t) = (\dot{T}_1(t), \dots, \dot{T}_K(t))$, where the ‘dot’ stands here for the right-derivative with respect to time, changes only on the event epochs σ_ℓ . Moreover, for $t \in [\sigma_\ell, \sigma_{\ell+1})$, $T_k(t) = 1$ if and only if $k = k^*$ where k^* is as in (7) and (8). Thus, $\dot{T}_k(t)$ is a measurable function with respect to the σ -algebra on \mathcal{X} and the Borel σ -algebra of $[0, 1]^K$. Since we generate service times only upon commencement of service the process Ξ is, under a queue-ratio discipline, a Markov process. Queue-ratio disciplines are a special case of head-of-the-line (HL) disciplines. We refer the reader to §3.1.5 of [38] for a formal construction of HL disciplines as Markov processes.

System dynamics Let $A_k(t)$, $k \in \mathcal{K}$, count the number of arrivals to class k by time t (both exogenous and from other classes). Let $D_k(t)$, $k \in \mathcal{K}$, count the number of service completions of class- k customers by time t and, for $j \in \mathcal{J}$, let $Y_j(t)$ be the cumulative idleness at station j by time t .

Throughout, the matrix Δ is fixed and we define the *nominal workload* $W = CMZ$. Note that in [38] and [5] W is used for the true immediate workload of which we do not keep track here. This abuse of notation facilitates making the needed connections to the antecedent literature. The process ϵ is then re-written as $\epsilon = Z - \Delta W$.

For each class k , we define $\epsilon_k^+(t) = \sum_{i: C_{s(k)}i=1, i>k} \epsilon_i(t)$. This is the “excess” at station $j = s(k)$ at time t corresponding to classes which are served in the same station as k but have higher indices. Also, we let $T_k^+(t)$ be the aggregate time allocated to these classes by time t . With these definitions, the dynamics of the network must satisfy the following equations for all $t \geq 0$,

$$A(t) = E(t) + \sum_k \varphi^k(D_k(t)), \quad (10)$$

$$Z(t) = Z(0) + A(t) - D(t), \quad (11)$$

$$\int_0^\infty W(t) dY(t) = 0, \quad (12)$$

$$Y(t) + CT(t) = et, \quad (13)$$

$$D(t) = S(T(t)), \quad (14)$$

$$t - T_k^+(t) \text{ can only increase when } \epsilon_k^+(t) = 0, \quad k = 1, \dots, K, \quad (15)$$

where the integral in (12) should be read componentwise and hereafter e denotes the J -dimensional vector with all elements equal to 1. Equation (14) holds for all HL disciplines. Equation (15) is equivalently written as

$$\int_0^\infty \epsilon_k^+(s) d(s - T_k^+(s)) = 0, \quad k = 1, \dots, K.$$

For the special case of preemptive resume static priority networks this reduces to the well known condition

$$\int_0^\infty Z_k^+(s) d(s - T_k^+(s)) = 0, \quad k = 1, \dots, K,$$

where Z_k^+ corresponds to the total queue of classes with higher priority than class k ; see e.g. [5, page 105].

2.3 Fluid model equations Three types of fluid models are used in the literature to specify sufficient conditions that guarantee edges (I-III) in the limit interchange diagram in Figure 1. In the context of proving SSC in [5] one considers fluid models that approximate the evolution of queueing network over short time intervals under a *hydrodynamic* scaling. The appropriately defined limits (*cluster points* in the

terminology of [5]) are expected to satisfy the following fluid-model equations:

$$\bar{Z}(t) = \bar{Z}(0) + \alpha t - (I - \tilde{P})M^{-1}\bar{T}(t) \geq 0, \quad (16)$$

$$\bar{W}(t) = CM\bar{Z}(t), \quad (17)$$

$$\bar{\epsilon}(t) = \bar{\epsilon}(0) + (I - \Delta CM)(\alpha t - (I - \tilde{P})M^{-1}\bar{T}(t)), \quad (18)$$

$$\bar{T}(t) \text{ is nondecreasing and starts from zero,} \quad (19)$$

$$\bar{Y}(t) = \alpha t - C\bar{T}(t) \text{ is nondecreasing,} \quad (20)$$

$$\int_0^\infty \bar{W}(s)d\bar{Y}(s) = 0, \quad (21)$$

$$\int_0^\infty \bar{\epsilon}_k^+(s)d(s - \bar{T}_k^+(s)) = 0, \quad k = 1, \dots, K. \quad (22)$$

These are natural deterministic counterparts of (10)-(15). The equation (18) is here redundant as it follows from (16) and the definition of $\bar{\epsilon}$ but it will be useful for the discussions that follow. All solutions to (16)-(22) are Lipschitz continuous and we let N be the Lipschitz constant (N is specified explicitly in §5.3). In the SSC framework of [5] one requires that solutions $\bar{X} = (\bar{W}, \bar{Z}, \bar{\epsilon}, \bar{T})$ to the fluid-model equations (16)-(22) are attracted to the invariant manifold. Towards limit interchange we strengthen this requirement to a linear drift requirement; see Theorem 3.1. In the following definition we say that $t \geq 0$ is regular for \bar{X} if $\dot{\bar{X}}(t)$ exists at t .

Definition 1 (piecewise-linear test functions for SSC) *A family of non-negative vectors $h = (h_1, \dots, h_n)$ is said to induce a piecewise-linear Lyapunov function for the fluid model if the following holds: There exists $c_0 > 0$ such that for any solution $\bar{X} = (\bar{W}, \bar{Z}, \bar{\epsilon}, \bar{T})$ to the fluid-model equations (16)-(22) and any regular $t \geq 0$,*

$$(a) \max_{1 \leq i \leq n} \langle h_i, \dot{\bar{\epsilon}}(t) \rangle = 0 \text{ if and only if } \|\bar{\epsilon}(t)\| = 0, \text{ and}$$

$$(b) \max_{1 \leq i \leq n} \langle h_i, \dot{\bar{\epsilon}}(t) \rangle \leq -c_0 \text{ if } \|\bar{\epsilon}(t)\| > 0.$$

If such a family of vectors exists we say that the fluid model is attracted to the invariant manifold at a linear rate.

The linear attraction to the invariant manifold guarantees, in particular, that once $\bar{\epsilon}$ is close to 0 it stays there *regardless* of, say, the specific value of \bar{Z} . As queue-ratio disciplines respond directly to the distance, ϵ , from the invariant manifold such linear attraction seems plausible. In the special case of static priorities Chen and Ye [10, Proposition 3.5] and Chen and Zhang [12, Theorem 4] identifies explicit algebraic conditions on the network primitives that guarantee that a piecewise-linear test function exists for the fluid model.

To have meaningful approximations under the hydrodynamic scaling one requires that the (sequence of) diffusion-scale queues at time $t = 0$ form a tight sequence (see e.g. Theorem 3 in [3]). To establish state-space collapse in steady-state, however, we will want to analyze the drift when the network is initialized with its steady-state distribution which we cannot assume a priori to be tight (as that is exactly what we seek to prove). It will suffice for our purposes to capture the increments of ϵ^r . In essence, we will consider limits under appropriate scaling of the (augmented and truncated process)

$$(W^r \wedge \Theta\sqrt{r}, Z^r \wedge \Theta\sqrt{r}, T^r, \epsilon^r \wedge \Theta\sqrt{r}, \mathcal{H}^r),$$

where $\mathcal{H}^r(t) = \epsilon^r(t) - \epsilon^r(0)$, Θ is a truncation constant and, for abbreviation, we write $W^r \wedge \Theta\sqrt{r} = (W_1^r \wedge \Theta\sqrt{r}, \dots, W_J^r \wedge \Theta\sqrt{r})$ and similarly for the other processes. Given a time interval $[0, L]$, we set $\Theta = 3NL$ where N is the Lipschitz constant of the fluid model equations (16)-(22).

The fluid model that emerges as an appropriate limit of these *truncated* scaled processes (see §5.3) is the following modification of (16)-(22):

$$(16') \quad \bar{Z}(t) \geq \bar{Z}(0) \wedge \Theta + \alpha t - (I - \tilde{P})M^{-1}\bar{T}(t) \geq 0,$$

$$(17') \quad \bar{W}(t) = CM\bar{Z}(t),$$

$$(18') \quad \bar{\epsilon}(t) \geq \bar{\epsilon}(0) \wedge \Theta + (I - \Delta CM)(\alpha t - (I - \tilde{P})M^{-1}\bar{T}(t)),$$

$$\begin{aligned}
 (18a') \quad & \bar{\mathcal{H}}(t) = (I - \Delta CM)(\alpha t - (I - \tilde{P})M^{-1}\bar{T}(t)), \\
 (19') \quad & \bar{T}(t) \text{ is nondecreasing and starts from zero,} \\
 (20') \quad & \bar{Y}(t) = et - C\bar{T}(t) \text{ is nondecreasing,} \\
 (21') \quad & \int_0^\infty \bar{W}(s)d\bar{Y}(s) = 0, \\
 (22') \quad & \int_0^\infty (\bar{\epsilon}_k^+(s) \wedge \Theta)d(s - \bar{T}_k^+(s)) = 0, \quad k = 1, \dots, K.
 \end{aligned}$$

We will refer to (16')-(22') as the *truncated-fluid-model equations*. Equations (16')-(18') and (19')-(22') are natural (truncated) counterparts of equations (16)-(22). In (18a'), $\bar{\mathcal{H}}$ is interpreted as capturing (in fluid scale) the increments of the process that tracks the distance from the invariant manifold $\bar{\epsilon}(t) - \bar{\epsilon}(0)$ (see below). We say that $\bar{X} = (\bar{W}, \bar{Z}, \bar{T}, \bar{\epsilon}, \bar{\mathcal{H}})$ solves the truncated fluid model if it satisfies equations (16')-(22') for all $t \leq L$ and if $\|\bar{X}(t_2) - \bar{X}(t_1)\| \leq N|t_2 - t_1|$ for all $0 \leq t_1 < t_2 \leq L$.

There is an evident link between the fluid-model equations and their truncated counterparts. Given a solution $(\bar{W}, \bar{Z}, \bar{\epsilon}, \bar{T})$ to the fluid model equations (16)-(22),

$$\bar{X}(t) = (\bar{W}(t), \bar{Z}(t), \bar{T}(t), \bar{\epsilon}(t), \bar{\epsilon}(t) - \bar{\epsilon}(0)),$$

where $\bar{\epsilon}(t) = \bar{Z}(t) - \Delta\bar{W}(t)$, is a solution to the truncated fluid model equations (16')-(22'). In particular, from every solution to the fluid-model equations we can construct a solution to the truncated fluid-model equations, but the converse is, in general, false because the truncated fluid model is under specified compared to the fluid model. It is useful, however, that the two models are equivalent in terms of their attraction to the invariant manifold.

We say that all solutions to the truncated fluid model equations are attracted to the invariant manifold if there exists $c_0 > 0$ such that, for any solution \bar{X} to the truncated fluid model equations items (a) and (b) of Definition 1 hold with $\bar{\epsilon}$ there replaced with $\bar{\mathcal{H}}$. The following lemma is proved in §F of the appendix.

Lemma 2.1 *If all solutions to the fluid model equations (16)-(22) are attracted to the invariant manifold at linear rate then so do all solutions to the truncated-fluid-model equations (16')-(22').*

Thus, whereas the attraction to the invariant manifold of the *truncated* fluid model plays a crucial role in our proofs, Lemma 2.1 allows us to state the sufficient conditions for limit interchange in terms of the better-understood fluid-model equations (16)-(22).

The discussion thus far will suffice for the statement of our main result in the next section which will be followed by further formalization of some of the key concepts. We end this section with some notational conventions that we use throughout the paper.

Additional notational conventions: For a Markov process $\Xi = (\Xi(t), t \geq 0)$ on a locally compact separable metric space \mathcal{X} we let \mathbb{P}_x be the probability distribution under which $\mathbb{P}\{\Xi(0) = x\} = 1$ for $x \in \mathcal{X}$ and $\mathbb{E}_x[\cdot] = \mathbb{E}[\cdot | \Xi(0) = x]$ be the expectation operator w.r.t. the probability distribution \mathbb{P}_x . Let \mathbb{P}_π denote the probability distribution under which $\Xi(0)$ is distributed according to π and put $\mathbb{E}_\pi[\cdot]$ to be the expectation operator w.r.t. this distribution. A probability distribution π defined on \mathcal{X} is said to be a stationary distribution if for every bounded continuous function f

$$\mathbb{E}_\pi[f(\Xi(t))] = \mathbb{E}_\pi[f(\Xi(0))], \text{ for all } t \geq 0.$$

It is said to be the steady-state distribution if for every such function and all $x \in \mathcal{X}$,

$$\mathbb{E}_x[f(\Xi(t))] \rightarrow \mathbb{E}_\pi[f(\Xi(0))] \text{ as } t \rightarrow \infty.$$

We let $\mathcal{C}^d[0, \infty)$ be the space of continuous functions from $[0, \infty)$ to \mathbb{R}^d . We let $\mathcal{D}^d = \mathcal{D}^d[0, \infty)$ be the space of all RCLL (Right Continuous with Left Limits) \mathbb{R}^d -valued functions, equipped with the Skorohod J_1 metric; see e.g. [37]. We use ' \Rightarrow ' to denote weak convergence as $r \rightarrow \infty$ with respect to this metric, and when discussing \mathbb{R}^d -valued random variables ' \Rightarrow ' will simply mean convergence in distribution as $r \rightarrow \infty$. For a vector-valued process $x \in \mathcal{D}^d[0, \infty)$, let $\|x\|_{s,T} = \sup_{s \leq t \leq T} \|x(t)\|$, where $\|x(t)\| = \sum_{k=1}^d |x_k(t)|$ and we remove the subscript s if $s = 0$. Finally, throughout, we use the term *absolute constant* to denote a finite and strictly positive constant that does not depend on the heavy-traffic index r (but that may depend on other parameters). We use c_0, c_1, \dots to denote such constants.

3. Statement and discussion of the main result To state our main result, we let $\widehat{Z}^r(t) = Z^r(rt)/\sqrt{r}$ be the diffusion scaled queue-length in the r^{th} network at time t and let $\widehat{W}^r = CM\widehat{Z}^r$. The (scaled) distance from the invariant manifold is then $\widehat{\epsilon}^r = \widehat{Z}^r - \Delta\widehat{W}^r$. Also, we let $\widehat{T}^r = T^r(rt)/\sqrt{r}$ be the diffusion scaled allocation process at time t , $\widehat{Y}^r(t) = Y^r(rt)/\sqrt{r}$ be the scaled cumulative idleness at time t and $\widehat{\mathcal{R}}_k^{a,r}(0) = \mathcal{R}_k^{a,r}(0)/\sqrt{r}$ and $\widehat{\mathcal{R}}_k^{v,r}(0) = \mathcal{R}_k^{v,r}(0)/\sqrt{r}$ be, respectively, the scaled residual inter-arrival and service times at time 0. Finally, α^r is the exogenous-arrival-rate vector in the r^{th} system. We assume that $\sqrt{r}(\alpha^r - \alpha) = \beta$ for some $\beta \in (-\infty, \infty)$ so that, in particular, $\alpha^r \rightarrow \alpha$ as $r \rightarrow \infty$; additional details regarding the scaling and the heavy-traffic conditions are provided in §4.

Some of the assumptions made in our main result below are borrowed directly from the literature and were shown to be sufficient for edges (I-III) in Figure 1. Specifically,

- (1) for the existence of the limit SRBM we impose certain structure on the data matrices, most importantly, that the reflection matrix $R = (CMQ\Delta)^{-1}$ satisfies a completely- \mathcal{S} condition. This is Assumption 7.1 in [38] which we will flesh out as Assumption 1 in §4.
- (2) for the positive recurrence of the SRBM we require that all solutions to a Skorohod problem (the “fluid model” of the SRBM) are attracted to the origin in finite time. This is the key assumption in Theorem 2.6 of [18] that we repeat here as Assumption 2 in §5.2.
- (3) for the positive recurrence of the queueing network we require that, for each index r along the sequence of networks, the corresponding fluid model is stable. This is the key assumption in Theorem 4.2 of [14] that we flesh out as Assumption 3 in §5.4;

When added to the above, the linear attraction to the invariant manifold, guarantees the validity of (IV) in the limit interchange diagram.

Theorem 3.1 (The main theorem) *Consider a sequence of queue-ratio networks in heavy-traffic and suppose that Assumptions 1, 2 and 3 hold and that any solution to the fluid model equations (16)-(22) is attracted to the invariant manifold at linear rate. We then have the following:*

- I. If $(\widehat{Z}^r(0), \widehat{\mathcal{R}}^{a,r}(0), \widehat{\mathcal{R}}^{v,r}(0)) \Rightarrow (Z(0), 0, 0)$, and $\epsilon^r \Rightarrow 0$ then $\widehat{Z}^r \Rightarrow \widehat{Z}$, where $\widehat{W} = CM\widehat{Z}$ is an SRBM.
- II. For all $r \in \mathbb{N}$, the process $\widehat{\Xi}^r$ has a unique stationary distribution which is also its steady-state distribution.
- III. The SRBM \widehat{W} has a unique stationary distribution which is also its steady-state distribution.
- IV. **Steady-state convergence:** *The sequence of steady-state queue-length vectors converges weakly*

$$\widehat{Z}^r(\infty) \Rightarrow \widehat{Z}(\infty),$$

where $CM\widehat{Z}(\infty)$ has the steady-state distribution of the SRBM \widehat{W} . Further, for any $m \in \mathbb{N}$,

$$\mathbb{E}[\|\widehat{Z}^r(\infty)\|^m] \rightarrow \mathbb{E}[\|\widehat{Z}(\infty)\|^m].$$

Discussion of the main result: On top of Assumptions 1-3 that follow previous literature, we impose in Theorem 3.1 two further requirements on queue-ratio networks. First, whereas the typical condition for state-space collapse is mere attraction to the invariant manifold, we require linear attraction. As mentioned above, for the special case of static priority networks, conditions that guarantee these have been explicitly related to algebraic properties of the underlying matrices but this remains to be verified for general queue-ratio disciplines. Second, the requirement that interarrival and service times have finite moments of all orders is an artifact of our proof techniques and it is plausible that this condition can be tightened. In fact, our proofs do not necessitate the existence of all such moments but we do require moments of significantly greater order than the mere second moment required in [38] and [5]. In our proofs we make explicit the dependence on p so as to underscore the sources of this requirement. The number of moments, m , for which the convergence in item (IV) of the theorem holds does depend on the value of p . However, the relation between the value of p in (4) and the number of moments, m , for which the convergence holds is not as clean as in the Generalized Jackson case (see [6]) where it was shown that such convergence holds for all $m < p - 1$.

3.1 Outline of the proof Here we provide an informal outline of the proof that highlights the key steps and ingredients for the proof of item (IV) in Theorem 3.1. Each step in this outline will be expanded upon and spelled out in detail in §5.

Step 1: Inclusion sets and Lyapunov functions Let $\widehat{\Xi} \equiv (\widehat{\Xi}(t), t \geq 0)$ be a continuous-time Markov process defined on a locally compact separable metric state space $\widehat{\mathcal{X}}$. For the special case of the queue-ratio networks, $\widehat{\Xi}$ would be the scaled version of (9). The following notion will be useful:

Definition 2 A function $\Phi : \widehat{\mathcal{X}} \rightarrow \mathbb{R}_+$ is said to be a constrained Lyapunov function of order $q \geq 1$ for $\widehat{\Xi}$ with drift-size parameter $-\delta < 0$, drift-time parameter $t_0 > 0$, exception parameter κ , and inclusion set $\mathcal{A} \subseteq \widehat{\mathcal{X}}$, if

$$\sup_{x \in \mathcal{A} : \Phi(x) > \kappa} \frac{\mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))] - \Phi^q(x)}{\Phi^{q-1}(x)} \leq -\delta. \quad (23)$$

The requirement that the initial state x belongs to the inclusion set \mathcal{A} is the distinguishing feature of constrained Lyapunov functions. In Proposition 5.1 we establish that, if $\Phi(\cdot)$ is a constrained Lyapunov function for a Markov process $\widehat{\Xi}$ that has a unique stationary distribution π , then under suitable conditions

$$\mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \right] \leq \left(1 + \frac{\varepsilon_1}{\delta} \right) \mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \mathbb{1}\{\widehat{\Xi}(0) \notin \mathcal{A}\} \right] + \frac{\varepsilon_2}{\delta}, \quad (24)$$

for constants $\varepsilon_1, \varepsilon_2 > 0$.

The introduction of constrained Lyapunov functions is motivated by particular characteristics of multi-class queueing networks in heavy traffic. Roughly speaking, as r increases, the queueing network exhibits state-space collapse and as a result “lives” in a small neighborhood of the invariant manifold. This neighborhood is expected to serve as an inclusion set for an appropriately chosen Lyapunov function.

Let $\widehat{\mathcal{X}}^r$ be the domain on which the process $\widehat{\Xi}^r$ takes its values; see §2.2. Given $r \in \mathbb{N}$ and $\epsilon > 0$, define

$$\mathcal{B}_\epsilon^r = \left\{ x = (z, \varrho^a, \varrho^v) \in \widehat{\mathcal{X}}^r : |\varrho^a| + |\varrho^v| \leq r^{-\epsilon}, k \in \mathcal{K} \right\}, \quad (25)$$

and

$$\mathcal{A}_\epsilon^r = \{ x \in \mathcal{B}_\epsilon^r : \|z - \Delta CMz\| \leq \epsilon \}. \quad (26)$$

In other words, \mathcal{A}_ϵ^r is the intersection of an ϵ -neighborhood of the invariant manifold with the family of states in which the (scaled) initial residuals are “well behaved.” A first step in the proof of Theorem 3.1 will be to identify a suitable constrained Lyapunov function $\Phi(\cdot)$ and show that the bound (24) holds for the diffusion-scaled queueing-network process $\widehat{\Xi}^r$ (see §4), with ε_1 there replaced by $c_0 r^{\frac{q}{2}}$ and with ε_2, δ not depending on r . We will then deduce tightness from (24) using properties of $\Phi(\cdot)$ and showing that

$$\limsup_{r \rightarrow \infty} r^{\frac{q}{2}} \mathbb{E}_{\pi^r} \left[\Phi^{q-1}(\widehat{\Xi}^r(0)) \mathbb{1}\{\widehat{\Xi}^r(0) \notin \mathcal{A}_\epsilon^r\} \right] < \infty. \quad (27)$$

Step 2: Identifying the constrained Lyapunov function Our point of departure here is the stability analysis of SRBM carried out by Dupuis and Williams [18] and summarized in Theorem 5.2 here. In that work, a (non-constrained) Lyapunov function $\Psi(\cdot)$ is used for the SRBM. Our constrained Lyapunov function is constructed from that function. Specifically, we will establish (see Proposition 5.2) that, for any constant $b > 1$, the function $\Phi(\cdot) = b + \Psi(\cdot)$ is a constrained Lyapunov function for the scaled queueing-network process $\widehat{\Xi}^r$, with inclusion set \mathcal{A}_ϵ^r as defined in (26). Intuitively, the “distance” between the queueing-network and its approximating SRBM is mostly captured by the distance of the queueing network from the invariant manifold. If the queueing network remains close to the invariant manifold, one expect that a “negative drift” for the SRBM (with the corresponding Lyapunov function) will translate into a similar drift for the queueing network. This logic assumes that, starting in the inclusion set, the network indeed remains close to the invariant manifold. This is the subject of step 3.

Step 3: Truncated fluid models and state-space collapse in steady-state To establish the concentration of π^r in \mathcal{A}_ϵ^r we will show in Theorem 5.3 that under the conditions of Theorem 3.1, the sequence of steady-state queues $\{\widehat{Z}^r(\infty), r \in \mathbb{N}\}$ satisfies $\widehat{Z}^r(\infty) - \Delta CM \widehat{Z}^r(\infty) \Rightarrow 0$. The truncated fluid model equations play a key role here. These allow us to prove SSC before (and independently of) proving the tightness of the scaled steady-state queues. We also show that, initializing the network in the inclusion set \mathcal{A}_ϵ^r , the network process remains close to the inclusion set; see Theorem 5.4. This is instrumental in establishing that the Lyapunov function, identified in step 2 above, is indeed a constrained Lyapunov function for the queueing network process.

Step 4: Crude steady-state bounds Establishing (27) requires moment bounds for the stationary queues and the starting point of this paper is that such bounds are not available a priori. Fortunately, if the probability $\mathbb{P}_{\pi^r}\{\Xi^r(0) \notin \mathcal{A}_\epsilon^r\}$ decays sufficiently fast, crude moment bounds are sufficient. To that end, we will show in Theorem 5.6 that for suitably large constants c_0, l and all $r \in \mathbb{N}$

$$\mathbb{E}_{\pi^r} \left[\|\widehat{Z}^r(0)\|^q \right] \leq c_0 r^l,$$

and, as a corollary, obtain (27) which will conclude the proof.

The remainder of the paper In §4 we define in detail the heavy-traffic scaling and review relevant diffusion-limits result from [38]. The main contribution of this paper is embedded in part IV of Theorem 3.1. This part is re-stated and proved in §5. Concluding remarks are provided in §6. Throughout, proofs of auxiliary results are relegated to the appendix.

4. The queueing network in heavy-traffic and diffusion limits

4.1 Heavy-traffic conditions and scaling We add the heavy-traffic index $r \in \mathbb{N}$ to all relevant processes and quantities defining the network to make the dependence on this index explicit, and omit it in the absence of such dependence.

The rate of exogenous arrivals to class k in the r^{th} network is denote by α_k^r (so that $\mathbb{E}[u_k^r(2)] = 1/\alpha_k^r$). The traffic equations for the r^{th} queueing network are then given by

$$\lambda^r = \alpha^r t + \widetilde{P} \lambda^r,$$

or equivalently by

$$\lambda^r = Q \alpha^r,$$

where λ_k^r , the k^{th} components of λ^r , denotes the total arrival rate for class k in the r^{th} system. We define the total traffic intensity ρ_j^r for the j^{th} station as

$$\rho_j^r = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k^r,$$

or in matrix form: $\rho^r = CMQ \alpha^r$ where $\rho^r = (\rho_1^r, \dots, \rho_J^r)'$.

Throughout we will assume that $M = \text{diag}(m_1, \dots, m_K)$, the coefficients of variation $c_{a,k}$, $k \in \mathcal{K}$ and $c_{s,k}$, $k \in \mathcal{K}$, as well as the routing matrix P , remain fixed and do not scale with r . This is assumed for simplicity of presentation and the analysis can be extended to the case in which these parameters are obtained as limits of corresponding sequences, M^r , P^r , $c_{a,k}^r$ and $c_{s,k}^r$ (see e.g. the analysis in [38] and [5]).

The sequence of systems defined above is said to be in *heavy-traffic* if $\alpha^r = \alpha + \beta/\sqrt{r}$ for some vector β , and a strictly positive vector α so that, for all $r \in \mathbb{N}$,

$$\sqrt{r}(1 - \rho_j^r) = \gamma_j, \tag{28}$$

for some vector $\gamma = (\gamma_1, \dots, \gamma_J)$. Since we restrict attention to cases in which the diffusion limit is stable, we will assume that γ has strictly positive entries.

Let

$$\bar{D}^r(t) = \frac{D^r(rt)}{r}, \quad \bar{T}^r(t) = \frac{T^r(rt)}{r},$$

denote the *fluid-scale* departure and time-allocation processes, respectively, and define the following *diffusion-scale* processes:

$$\widehat{E}^r(t) = \frac{E^r(rt) - \alpha^r rt}{\sqrt{r}}, \quad \widehat{\varphi}^{k,r}(t) = \frac{\varphi^{k,r}(\lfloor rt \rfloor) - \widetilde{P}^k \lfloor rt \rfloor}{\sqrt{r}}, \quad \widehat{D}^r(t) = \frac{D^r(rt) - \mu rt}{\sqrt{r}},$$

and

$$\widehat{Z}^r(t) = \frac{Z^r(rt)}{\sqrt{r}}, \quad \widehat{Y}^r(t) = \frac{Y^r(rt)}{\sqrt{r}}, \quad \text{and} \quad \widehat{W}^r(t) = CM\widehat{Z}^r.$$

We also write

$$\widehat{\mathcal{R}}^{a,r}(t) = \frac{R^{a,r}(rt)}{\sqrt{r}}, \quad \widehat{\mathcal{R}}^{v,r}(t) = \frac{R^{v,r}(rt)}{\sqrt{r}}.$$

Finally, we define the scaled version of (9)

$$\widehat{\Xi}^r = (\widehat{Z}^r, \widehat{\mathcal{R}}^{a,r}, \widehat{\mathcal{R}}^{v,r}), \tag{29}$$

and let $\widehat{\mathcal{X}}^r$ denote its domain.

We write

$$R = (CMQ\Delta)^{-1}, \tag{30}$$

$$\widehat{\epsilon}^r(t) = \widehat{Z}^r(t) - \Delta\widehat{W}^r(t), \tag{31}$$

$$\widehat{\eta}^r(t) = CMQ\widetilde{P}(\widehat{\epsilon}^r(0) - \widehat{\epsilon}^r(t)), \tag{32}$$

$$\widehat{\xi}^r(t) = -CM\widehat{S}^r(\bar{T}^r(t)) + CMQ \left(\widehat{E}^r(t) + \sum_{k=1}^K \widehat{\varphi}^{k,r}(\bar{D}_k^r(t)) \right) - \gamma t, \tag{33}$$

$$\widehat{X}^r(t) = \widehat{W}^r(0) + R(\widehat{\xi}^r(t) + \widehat{\eta}^r(t)), \tag{34}$$

so that by (10)-(14),

$$\widehat{W}^r(t) = \widehat{X}^r(t) + R\widehat{Y}^r(t). \tag{35}$$

The lifting matrix Δ in (30) and (31) is as in (6).

Put

$$H = C \left(\Lambda\Sigma + MQ \left(\Pi + \sum_{k=1}^K \lambda_k \Upsilon^k \right) Q'M \right) C', \tag{36}$$

where

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_K), \quad \Pi = \text{diag}(\alpha_1 c_{a,1}^2, \dots, \alpha_K c_{a,K}^2), \\ \Sigma = \text{diag}(m_1^2 c_{s,1}^2, \dots, m_K^2 c_{s,K}^2),$$

and Υ is defined in (5). The above construction implicitly presumes, then, that the matrix $CMQ\Delta$ is invertible. This is formally stated in Assumption 7.1 of [38] that we repeat below and for which it is said that a $J \times J$ matrix R is completely- \mathcal{S} , if and only if for each principal submatrix \check{R} of R , there is a vector $\nu > 0$ such that $\check{R}\nu > 0$.

Assumption 1 (Data Matrices) (i) *The matrix $CMQ\Delta$ is invertible and $R = (CMQ\Delta)^{-1}$ is completely- \mathcal{S} .* (ii) *The matrix H given in (36) is strictly positive definite.*

Assumption 1 completes the description of the system parameters, dynamics and scaled processes.

4.2 The Brownian system model The “natural” diffusion analogue of the queueing network is captured mathematically by means of a semi-martingale reflecting Brownian motion (SRBM). The definition of SRBM with data $(S, \theta, \Gamma, R, \nu)$ is given below (we refer to §6 of [38] for further discussion of the SRBM and relevant references). Throughout this section we fix $S = \mathbb{R}_+^J$ and a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$. Let \mathcal{B} be the σ -algebra of Borel subsets of S . Let θ be a constant vector in \mathbb{R}^J , Γ a $J \times J$ non-degenerate covariance matrix (symmetric and strictly positive definite), and R a $J \times J$ matrix.

Definition 3 (SRBM) Given a probability measure ν on (S, \mathcal{B}) , an SRBM associated with the data $(S, \theta, \Gamma, R, \nu)$ is an \mathcal{F}_t -adapted, J -dimensional process W such that

- (i) $W = X + RY$, \mathbb{P} -a.s.,
- (ii) \mathbb{P} -a.s., W has continuous paths and $W(t) \in S$ for all $t \geq 0$,
- (ii) under \mathbb{P} ,
 - (a) X is a J -dimensional Brownian motion with drift vector θ , covariance matrix Γ and $X(0)$ has distribution ν ,
 - (b) $(\{X(t) - X(0) - \theta t, \mathcal{F}_t, t \geq 0\})$ is a martingale,
- (iv) Y is an \mathcal{F}_t -adapted, J -dimensional process such that \mathbb{P} -a.s. for each $j \in \mathcal{J}$,
 - (a) $Y_j(0) = 0$,
 - (b) Y_j is continuous and nondecreasing,
 - (c) Y_j can increase only when W is on the face $F_j \equiv \{x \in S : x_j = 0\}$, i.e.,

$$\int_0^\infty W_j(s) dY_j(s) = 0.$$

When discussing steady state, the initial distribution ν is immaterial and we will refer to the SRBM with data (S, θ, Γ, R) . The following is an adaptation of the main result in Williams [38].

Theorem 4.1 (II: Diffusion limits) Suppose that Assumption 1 holds and that

$$\left(\widehat{Z}^r(0), \widehat{R}^{a,r}(0), \widehat{R}^{v,r}(0) \right) \Rightarrow (\widehat{Z}(0), 0, 0),$$

where $\widehat{W}(0) = CM\widehat{Z}(0)$ has distribution ν . Suppose further that state-space collapse holds, i.e., that

$$\widehat{\epsilon}^r \Rightarrow 0.$$

Then,

$$\widehat{Z}^r \Rightarrow \Delta \widehat{W},$$

where \widehat{W} is an SRBM associated with the data $(S, \theta, \Gamma, R, \nu)$ for $\Gamma = RHR'$ and $\theta = -R\gamma$.

Recall that our construction of the queueing network is different than that of [38] in that we generate the customer service times only upon entrance to service, rather than upon arrival of a customer to a station. For queue-ratio disciplines our construction is, however, equivalent to that of [38] in that, starting empty, the process Z^r has the same probability law under both constructions and, in turn, both constructions will share the same diffusion limits. This equivalence holds also if both constructions are initialized at time 0 with the same distribution of residuals and provided that, in [38], the service times of the customers in queue at time 0 are i.i.d. and distributed according to $F_k^s(\cdot)$. Thus, Theorem 4.1 is a direct corollary of Theorem 7.1 of [38].

Theorem 4.1 hints to the applicability of queue-ratio disciplines. In the more general result of [38], given state-space collapse, the law of the diffusion limit is determined by the initial distribution ν , the data matrices R, Γ, Δ and the vector γ . (R itself is also defined through Δ). In turn, since a queue-ratio discipline can be defined for arbitrary lifting matrices Δ as in (6), it stands to reason that, asymptotically, any law for the queue length vector that is covered by the general results of [38] can be achieved via the corresponding queue-ratio discipline provided. The formalization of this statement is beyond the scope of this paper; see further discussion in §6.

5. Re-statement of the main result and completion of the proof Our main contribution is concerned with the steady-state approximation as embedded in statement IV of Theorem 3.1 which we now restate and prove.

Theorem 5.1 (IV: Steady-state convergence) Under the conditions of Theorem 3.1, it holds that

$$\widehat{Z}^r(\infty) \Rightarrow \Delta \widehat{W}(\infty),$$

where $\widehat{W}(\infty)$ has the steady-state distribution of the SRBM with data $(S, -R\gamma, \Gamma, R)$.

We prove Theorem 5.1 by elaborating on the outline provided in §3. Sections 5.1, 5.2, 5.3 and 5.4 are dedicated, respectively, to steps 1-4 in that outline. Section 5.5 combines all the steps to conclude the proof of this theorem.

5.1 Inclusion sets and Lyapunov functions Given a Markov process $\widehat{\Xi} = (\widehat{\Xi}(t), t \geq 0)$ on a locally compact separable metric state space $\widehat{\mathcal{X}}$, a subset $\mathcal{A} \subseteq \widehat{\mathcal{X}}$ and a function $\Phi(\cdot) : \widehat{\mathcal{X}} \rightarrow \mathbb{R}_+$ we define for all $q \in \mathbb{N}$,

$$\phi_q^{\widehat{\Xi}}(t, \mathcal{A}) = \sup_{x \in \mathcal{A}} \Phi^{-(q-1)}(x) \mathbb{E}_x \left[(\Phi^q(\widehat{\Xi}(t)) - \Phi^q(x))^+ \right], \quad (37)$$

where the expectation may be infinite. Below, the notion of constrained Lyapunov function is as in Definition 2.

Proposition 5.1 *Suppose that the Markov process $\widehat{\Xi}$ possesses a stationary distribution π . Assume that Φ is a constrained Lyapunov function of order $q \geq 1$ with drift-size parameter $-\delta < 0$, drift-time parameter $t_0 > 0$, exception parameter κ and inclusion set $\mathcal{A} \subseteq \widehat{\mathcal{X}}$, such that:*

- (a) $\phi_q^{\widehat{\Xi}}(t_0, \mathcal{A})$ is finite;
- (b) $\mathbb{E}_\pi[\Phi^q(\widehat{\Xi}(0))] < \infty$;
- (c) $\mathbb{E}_\pi \left[(\Phi^q(\widehat{\Xi}(t_0)) - \Phi^q(\widehat{\Xi}(0))) \mathbb{1}\{\widehat{\Xi}(0) \notin \mathcal{A}\} \right] \leq \varepsilon_0 \mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \mathbb{1}\{\widehat{\Xi}(0) \notin \mathcal{A}\} \right]$ for some constant $\varepsilon_0 > 0$.

Then,

$$\mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \right] \leq \left(1 + \frac{\varepsilon_0}{\delta} \right) \mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \mathbb{1}\{\widehat{\Xi}(0) \notin \mathcal{A}\} \right] + \frac{\kappa^{q-1} \phi_q^{\widehat{\Xi}}(t_0, \mathcal{A})}{\delta}. \quad (38)$$

If, in addition, there exists ε_1 such that $\mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \mathbb{1}\{\widehat{\Xi}(0) \notin \mathcal{A}\} \right] \leq \varepsilon_1$, then

$$\mathbb{P}_\pi \{ \Phi^{q-1}(\widehat{\Xi}(0)) > y \} \leq \frac{\varepsilon_1}{y} \left(1 + \frac{\varepsilon_0}{\delta} \right) + \frac{\kappa^{q-1} \phi_q^{\widehat{\Xi}}(t_0, \mathcal{A})}{\delta y}. \quad (39)$$

5.2 Identifying the queueing-network constrained Lyapunov function The “fluid model” of the SRBM is informally obtained by removing the Brownian term to get a Skorohod Problem (SP) as defined below and where, as before, $S = \mathbb{R}_+^J$.

Definition 4 (Skorohod problem (SP)) *A pair $(\phi, \eta) \in \mathcal{C}^J[0, \infty) \times \mathcal{C}^J[0, \infty)$ solves the SP with respect to (S, θ, R, x) if the following holds:*

- (i) $\phi(t) = x + \theta t + R\eta(t) \in S$, for all $t \geq 0$;
- (ii) η is such that, for $i = 1, \dots, J$,
 - (a) $\eta_i(0) = 0$,
 - (b) η_i is nondecreasing and
 - (c) $\int_0^t \mathbb{1}\{\phi_i(s) \neq 0\} d\eta_i(s) = 0$ for all $t \geq 0$.

A solution (ϕ, η) is said to be attracted to the origin in finite time if for any $\epsilon > 0$ there exists $t_\epsilon < \infty$ such that $|\phi(t)| \leq \epsilon$ for all $t \geq t_\epsilon$.

The following is the main assumption made in [18] for purposes of stability of the SRBM.

Assumption 2 *Assumption 1 holds and for any initial state x , the ϕ component of all solutions to the SP with data (S, θ, R, x) is attracted to the origin in finite time.*

Resolving the question of attraction to the origin is not a trivial task (see e.g. [8]), but is not a focal point for the present paper. For our purposes, the following result is pertinent.

Theorem 5.2 (III: Stability of the SRBM - Theorem 2.6 in [18] and Theorem 4.12 in [7]) *Suppose that Assumption 1 holds and that for any initial state x , the ϕ component of all solutions of the SP with data (S, θ, R, x) is attracted to the origin in finite time. Then, the SRBM with data $(S, \theta, \Gamma, R, \nu)$ is positive recurrent and has a unique stationary distribution which is also its steady-state distribution.*

In the process of proving Theorem 5.2, Dupuis and Williams established the existence of a Lyapunov function $\Psi(\cdot) : S \rightarrow \mathbb{R}_+$ for the SRBM and proved that it satisfies some properties that will be useful for our analysis.

(P1) $\Psi(\cdot) \in C^2(S \setminus \{0\})$.

(P2) Given $\mathcal{N} < \infty$, there exists $\mathcal{W} < \infty$ such that $\Psi(w) \geq \mathcal{N}$ for all $w \in S$ with $\|w\| \geq \mathcal{W}$.

(P3) Given $\epsilon > 0$, there exists $\mathcal{W} < \infty$ such that $\|D^2\Psi(w)\| \leq \epsilon$ for all $w \in S$ with $\|w\| \geq \mathcal{W}$.

(P4) There exists $\epsilon_0 > 0$ such that

$$\begin{aligned} D\Psi(w) \cdot \theta &\leq -\epsilon_0, & \text{for all } w \in S \setminus \{0\}, \\ D\Psi(w) \cdot y &\leq -\epsilon_0, & \text{for all } y \in y(w), w \in \partial S \setminus \{0\}, \end{aligned} \quad (40)$$

where

$$y(w) = \left\{ \sum_{j=1}^J q_j R^j : \sum_{j=1}^J q_j = 1, q_j \geq 0, \text{ and } q_j > 0 \text{ only if } w_j = 0 \right\}.$$

Here R^j is the j^{th} column of the matrix R defined in (30).

(P5) $\Psi(\cdot)$ is radially homogeneous: $\Psi(\alpha w) = \alpha\Psi(w)$ for $\alpha \geq 0$, $w \in S$.

(P6) $\varpi = \sup_{w \in S \setminus \{0\}} \|D\Psi(w)\| < \infty$.

(P7) There exist $\epsilon_1, \epsilon_2 \in (0, \infty)$ such that $\epsilon_1\|w\| \leq \Psi(w) \leq \epsilon_2\|w\|$, for all $w \in \mathbb{R}_+^J$.

Properties (P6) and (P7) are derived in Theorem 4.1 of Budhiraja and Lee [7].

Fix a constant $b > 1$ and define a mapping $\Phi(\cdot) : \widehat{\mathcal{X}}^r \rightarrow \mathbb{R}_+$ by letting, for $x = (z, \varrho^a, \varrho^v) \in \widehat{\mathcal{X}}^r$,

$$\Phi(x) = b + \Psi(CMz). \quad (41)$$

Below, \mathcal{A}_ϵ^r is as in (26), $\widehat{\Xi}^r$ is the scaled network process as in (29) and $\phi_q^{\widehat{\Xi}^r}(\cdot, \cdot)$ is as in (37).

Proposition 5.2 (the constrained Lyapunov function) *Suppose that the conditions of Theorem 3.1 hold and fix $\epsilon > 0$ and $q \geq 1$. Then, there exist absolute constants δ , t_0 and κ such that, for all sufficiently large r , $\Phi(\cdot)$ is a constrained Lyapunov function of order q for $\widehat{\Xi}^r$ with drift-size parameter $-\delta < 0$, drift-time parameter t_0 , exception parameter κ , and inclusion set \mathcal{A}_ϵ^r . Moreover,*

$$\kappa' := \limsup_{r \rightarrow \infty} \phi_q^{\widehat{\Xi}^r}(t_0, \mathcal{A}_\epsilon^r) < \infty. \quad (42)$$

Proposition 5.2 plays a key role in the proof of our main result. Its proof appears at the end of this section.

5.3 State-space collapse via truncated fluid models Recall that, fixing $L > 0$, we say that $\bar{X} = (\bar{W}, \bar{Z}, \bar{\epsilon}, \bar{T})$ solves the truncated fluid model equations on $[0, L]$ if it satisfies equations (16)-(22) for all $t \leq L$. Below we say that the process $\widehat{\Xi}^r$ is stable if it is positive Harris recurrent (see §3 of [14]). Also, the sets \mathcal{B}_ϵ^r and \mathcal{A}_ϵ^r are as defined in (25) and (26).

Theorem 5.3 (SSC in stationarity) *Suppose that, for each $r \in \mathbb{N}$, the process $\widehat{\Xi}^r$ is stable and let π^r be the corresponding (unique) stationary distribution. Suppose further that the fluid model (16)-(22) is attracted to the invariant manifold at linear rate. Then, given $\epsilon, T > 0$ and $m \in \mathbb{N}$, there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,*

$$\mathbb{P}_{\pi^r} \{ \|\widehat{\epsilon}^r\|_T > \epsilon \} \leq \varepsilon r^{-m}, \quad (43)$$

and

$$\mathbb{P}_{\pi^r} \{ \widehat{\Xi}^r(0) \notin \mathcal{B}_\epsilon^r \} \leq \varepsilon r^{-m}. \quad (44)$$

In turn, if $\widehat{\Xi}^r$ is initialized at time 0 with its stationary distribution, then,

$$\left(\widehat{\mathcal{R}}^{a,r}(0), \widehat{\mathcal{R}}^{v,r}(0)\right) \Rightarrow (0, 0) \text{ and } \widehat{\epsilon}^r \Rightarrow 0. \quad (45)$$

By Lemma 2.1 the linear attraction of the fluid model (16)-(22) to the invariant manifold implies the same for the truncated fluid model. This linear attraction lies at the core of the proof of Theorem 5.3. It is the truncated fluid models that allow us to prove this result before (and in independently of) proving the tightness of the stationary queue lengths.

The next theorem is instrumental in establishing that $\Phi(\cdot)$ is indeed a constrained Lyapunov function for the queueing network; see Definition 2 and Proposition 5.2. Informally, the theorem states that, initialized in the small neighborhood of the invariant manifold, the process $\widehat{\Xi}^r$ will stay there.

Theorem 5.4 (probability bounds for SSC) Fix $\epsilon, T > 0$ and, $q, m \in \mathbb{N}$. Assume that the fluid model (16)-(22) is attracted to the invariant manifold at linear rate. Then, there exists an absolute constant ϵ (not depending on ϵ) such that, for all sufficiently large $r \in \mathbb{N}$,

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{P}_x \{ \|\widehat{\epsilon}^r\|_T > \epsilon \} \leq \epsilon r^{-m}, \quad (46)$$

and, for $0 < s \leq T$,

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{P}_x \{ \|\widehat{\epsilon}^r\|_{s,T} > \epsilon \} \leq \epsilon r^{-m}. \quad (47)$$

Finally,

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\widehat{\epsilon}^r\|_T^q] \leq \epsilon \epsilon. \quad (48)$$

In particular, if $(\widehat{\epsilon}^r(0), r^\epsilon \widehat{\mathcal{R}}^{a,r}(0), r^\epsilon \widehat{\mathcal{R}}^{v,r}(0)) \Rightarrow (0, 0, 0)$, then $\widehat{\epsilon}^r \Rightarrow 0$.

5.4 Crude steady-state bounds Step 4 of the outline in §3 is concerned with a crude bound on the steady-state queue length which we state formally in this section. Our starting point is Dai's [14] result that relates the stability of a fluid model to that of the underlying queueing network. This fluid model is obtained, for fixed r , by letting the initial conditions grow and using a proper scaling; see §4 of [14]. Any limit point then satisfies the following system of equations:

$$\bar{Z}^r(t) = \bar{Z}^r(0) + \alpha^r t - (I - \tilde{P})M^{-1}\bar{T}^r(t), \quad (49)$$

$$\bar{W}^r(t) = CM\bar{Z}^r(t), \quad (50)$$

$$\bar{\epsilon}^r(t) = \bar{\epsilon}^r(0) + (I - \Delta CM)(\alpha^r t - (I - \tilde{P})M^{-1}\bar{T}^r(t)), \quad (51)$$

$$\bar{T}^r(t) \text{ is nondecreasing and starts from zero,} \quad (52)$$

$$\bar{Y}^r(t) = et - C\bar{T}^r(t) \text{ is nondecreasing,} \quad (53)$$

$$\int_0^\infty \bar{W}^r(s) d\bar{Y}^r(s) = 0, \quad (54)$$

$$\int_0^\infty (\bar{\epsilon}_k^r)^+(s) d(s - (\bar{T}_k^r)^+(s)) = 0, \quad k = 1, \dots, K. \quad (55)$$

To contrast these equations with (16)-(22) we refer to these as the r^{th} fluid model equations. With the exception of the explicit dependence on r through α^r equations (49)-(55) are identical to (16)-(22).

The next assumption and the theorem that follows are from [14] and [16].

Assumption 3 For each r , the r^{th} fluid model is stable: there exists a fixed time t_0 such that for any solution to the r^{th} fluid model equations with $\|\bar{Z}^r(0)\| = 1$ it holds that $\bar{Z}^r(t) = 0$ for all $t \geq t_0$.

For the following recall that (4) is assumed to hold for all $p \in \mathbb{N}$.

Theorem 5.5 (I: stability for fixed r – Theorem 4.2 of [14] and Theorem 4.1 of [16]) Fix $r \in \mathbb{N}$. Suppose that Assumption 3 holds. Then, the Markov process $\widehat{\Xi}^r$ is positive Harris recurrent and has a unique stationary distribution π^r , which is also its steady-state distribution. Further, for any $q \in \mathbb{N}$,

$$\mathbb{E}_{\pi^r} [\|\widehat{Z}^r(0)\|^q] < \infty.$$

The fluid model equations (49)-(55) together with the analysis framework used in [16] allow us to obtain the following result which is, as explained in §3, instrumental in the proof of our main result.

Theorem 5.6 *Suppose that the conditions of Theorem 3.1 hold, fix $q \in \mathbb{N}$, and let π^r be the steady-state distribution of the Markov process $\widehat{\Xi}^r$. Then, there exist absolute constants ε_q and $n_q \in \mathbb{N}$ such that, for all $r \in \mathbb{N}$,*

$$\mathbb{E}_{\pi^r} \left[\|\widehat{Z}^r(0)\|^q \right] \leq \varepsilon_q r^{n_q}. \quad (56)$$

Remark 5.7 (on the proof of Theorem 5.6 and its relation to [6]) Theorem 5.6 is proved in §C of the appendix. The arguments are reminiscent of those used in [6] to establish tightness of the diffusion-scale steady-state queues in the case of the single class generalized Jackson networks. As in [6], our proof of Theorem 5.6 is based on a double scaling (in r and in the initial condition) and on certain parts of the analysis in [16]. For Jackson networks, the continuity of the corresponding Skorohod problem is used by [6] to yield the tightness of the diffusion-scale steady-state queues. In the multiclass setting, in which such a continuity is absent, the double-scaling allows us to obtain a cruder bound. Within our framework, this crude bound is sufficient towards establishing the diffusion-scale tightness as outlined in §3.1.

The main challenge in the proof of Theorem 5.6 is in capturing the dependence on (the heavy-traffic index) r of the decay-rate of the fluid model towards the origin. This is achieved by first establishing a state-space-collapse result under the appropriate double scaling and, subsequently, relating the (doubly) scaled network dynamics to a Skorohod problem whose decay rate to 0 does not depend on r . Thus, the proof combines the attraction of the SP to the origin (Assumption 2) with the attraction of the fluid model (16)-(22) to the invariant manifold, to obtain an initial bound on the steady-state queues. ■

5.5 Completing the proof of Theorem 5.1

Proof of Theorem 5.1: We first verify that the conditions of Proposition 5.1 hold with respect to the function $\Phi(\cdot)$ and the Markov process $\widehat{\Xi}^r$. This will allow us to use (38) towards establishing tightness.

First, by Proposition 5.2, $\Phi(\cdot)$ is a constrained Lyapunov function for $\widehat{\Xi}^r$ with inclusion set \mathcal{A}_ϵ^r and $\kappa' := \limsup_r \phi_q^{\widehat{\Xi}^r}(t_0, \mathcal{A}_\epsilon^r) < \infty$. In particular, condition (a) in Proposition 5.1 is satisfied for all sufficiently large r . Let π^r be the stationary distribution of the Markov process $\widehat{\Xi}^r$ (see Theorem 5.5). Using crude bounds on the arrival process we obtain

$$\mathbb{E}_{\pi^r} \left[(\Phi^q(\widehat{\Xi}^r(t_0)) - \Phi^q(\widehat{\Xi}^r(0))) \mathbb{1}\{\widehat{\Xi}^r(0) \notin \mathcal{A}_\epsilon^r\} \right] \leq c_0 r^{\frac{q}{2}} \mathbb{E}_{\pi^r} \left[\Phi^{q-1}(\widehat{\Xi}^r(0)) \mathbb{1}\{\widehat{\Xi}^r(0) \notin \mathcal{A}_\epsilon^r\} \right]. \quad (57)$$

for all $r \in \mathbb{N}$. The simple proof of this bound appears in §F of the appendix. Thus, condition (b) of Proposition 5.1 is satisfied. Theorem 5.5 guarantees that so is condition (c). Then,

$$\mathbb{E}_{\pi^r} [\Phi^{q-1}(\widehat{\Xi}^r(0))] \leq \left(1 + \frac{c_0 r^{\frac{q}{2}}}{\delta} \right) \mathbb{E}_{\pi^r} \left[\Phi^{q-1}(\widehat{\Xi}^r(0)) \mathbb{1}\{\widehat{\Xi}^r(0) \notin \mathcal{A}_\epsilon^r\} \right] + \frac{\kappa^{q-1} \phi_q^{\widehat{\Xi}^r}(t_0, \mathcal{A}_\epsilon^r)}{\delta}, \quad (58)$$

and

$$\mathbb{P}_{\pi^r} \{ \Phi^{q-1}(\widehat{\Xi}^r(0)) > y \} \leq \frac{1}{y} \left(1 + \frac{c_0 r^{\frac{q}{2}}}{\delta} \right) \mathbb{E}_{\pi^r} \left[\Phi^{q-1}(\widehat{\Xi}^r(0)) \mathbb{1}\{\widehat{\Xi}^r(0) \notin \mathcal{A}_\epsilon^r\} \right] + \frac{\kappa^{q-1} \phi_q^{\widehat{\Xi}^r}(t_0, \mathcal{A}_\epsilon^r)}{\delta y}. \quad (59)$$

Recalling that, for $x \in \widehat{\mathcal{X}}^r$, $\Phi(x) = b + \Psi(CMz)$ and using property (P7) of the function $\Psi(\cdot)$ we have

$$b + c_1 \|CM\widehat{Z}^r(0)\| \leq \Phi(\widehat{\Xi}^r(0)) \leq b + c_2 \|CM\widehat{Z}^r(0)\|. \quad (60)$$

Let $n_{2(q-1)}$ be as in Theorem 5.6 and choose m in Theorem 5.3 so that $m - n_{2(q-1)} > q$. Applying Hölder's inequality we have that

$$\limsup_{r \rightarrow \infty} r^{\frac{q}{2}} \mathbb{E}_{\pi^r} \left[\Phi^{q-1}(\widehat{\Xi}^r(0)) \mathbb{1}\{\widehat{\Xi}^r(0) \notin \mathcal{A}_\epsilon^r\} \right] \leq c_3, \quad (61)$$

and it follows from (59) that

$$\lim_{y \rightarrow \infty} \limsup_r \mathbb{P}_{\pi^r} \{ \Phi^{q-1}(\widehat{\Xi}^r(0)) > y \} = 0.$$

In particular the sequence $\{ \Phi^{q-1}(\widehat{\Xi}^r(0)), r \in \mathbb{N} \}$ is tight and, since $\|CMz\| \geq c_4\|z\|$ for all $z \in \mathbb{R}_+^K$, so is the sequence $\{\widehat{Z}^r(0), r \in \mathbb{N}\}$.

The convergence now follows from tightness through a standard argument. Consider the sequence of queueing networks where each element in the sequence is initialized at time $t = 0$ with its stationary distribution. Since the sequence $\{\widehat{Z}^r(0), r \in \mathbb{N}\}$ is tight, every subsequence $\{\widehat{Z}^{r_j}(0), j \geq 1\}$ contains a convergent subsequence. Fix such a convergent subsequence $\{\widehat{Z}^{r_{j_l}}(0), l \geq 1\}$ and let $\widehat{Z}(0)$ be its weak limit. The convergence $\widehat{Z}^{r_{j_l}}(0) \Rightarrow \widehat{Z}(0)$, together with (45) allows us to apply Theorem 4.1 to conclude that $\widehat{Z}^{r_{j_l}} \Rightarrow \widehat{Z}$ where $\widehat{W} = CM\widehat{Z}$ is an SRBM with data $(S, -R\gamma, \Gamma, R, \nu)$ and ν is the distribution of $CM\widehat{Z}(0)$. As we initialized the process $\widehat{\Xi}^{r_{j_l}}$ with a stationary distribution we have that $\widehat{Z}^{r_{j_l}}(t) \stackrel{d}{=} \widehat{Z}^{r_{j_l}}(0)$ for all $t \geq 0$, and in particular, $\widehat{Z}^{r_{j_l}}(t) \Rightarrow \widehat{Z}(0)$ for all such t . In turn, $\widehat{Z}(t) \stackrel{d}{=} \widehat{Z}(0)$ for all $t \geq 0$ so that $CM\widehat{Z}(t)$ must be distributed according to a stationary distribution of the SRBM. As this distribution is unique (Theorem 5.2), $CM\widehat{Z}(0)$ must have that distribution. These arguments apply to any convergent subsequence and we conclude that $\widehat{Z}^r(0) \Rightarrow \widehat{Z}(0)$ where $CM\widehat{Z}(0)$ has the steady-state distribution of the SRBM. Finally, given $m \in \mathbb{N}$, set $q - 1 > m$ in (58) and (61) to conclude that the sequence $\|\widehat{Z}^r(\infty)\|^m$ is uniformly integrable so that the convergence of the expectations follows. ■

We conclude this section with the proof of Proposition 5.2. We require several auxiliary results. The first, Proposition (5.3) allows us to construct a constrained Lyapunov function of order $q > 1$ from one of order $q = 1$; see Definition 2. To that end, for a Markov process $\widehat{\Xi}$ on a locally compact separable metric state space $\widehat{\mathcal{X}}$ define,

$$L_q^{\widehat{\Xi}}(t, \mathcal{A}) = \sup_{x \in \mathcal{A}} \Phi^{-(q-2)}(x) \mathbb{E}_x \left[(\Phi(\widehat{\Xi}(t)) - \Phi(x))^2 (\Phi(x) + |\Phi(\widehat{\Xi}(t)) - \Phi(x)|)^{q-2} \right], \quad (62)$$

where the expectation may be infinite. Below $\phi_q^{\widehat{\Xi}}(\cdot, \cdot)$ is as in (37).

Proposition 5.3 *Suppose that Φ is a constrained Lyapunov function of order 1 for $\widehat{\Xi}$ with parameters k, t_0, κ and inclusion set \mathcal{A} and fix $q > 1$. Then, provided that $L_q^{\widehat{\Xi}}(t_0, \mathcal{A})$ and $\phi_q^{\widehat{\Xi}}(t_0, \mathcal{A})$ are finite, $\Phi(\cdot)$ is a constrained Lyapunov function of order q with drift-size parameter $-\delta q/2$, drift-time parameter t_0 , exception parameter $\max\{\kappa, L_q^{\widehat{\Xi}}(t_0, \mathcal{A})(q-1)/\delta\}$ and inclusion set \mathcal{A} .*

Next, recall that

$$\widehat{W}^r(t) = \widehat{W}^r(0) + R(\widehat{\xi}^r(t) + \widehat{\eta}^r(t)) + R\widehat{Y}^r(t),$$

with the processes $\widehat{\eta}^r$ and $\widehat{\xi}^r$ as defined in (32) and (33). Below, for a process $x \in \mathcal{D}^d[0, \infty)$ and a time $t > 0$, we let $\Delta x(t) = x(t) - x(t-)$. This should not be confused with the lifting matrix.

Lemma 5.8 below is an adaptation of Lemmas 8.3 and 8.4 of [38] with modifications to our simpler setting.

Lemma 5.8 *Fix $\epsilon > 0$, $r \in \mathbb{N}$ and an initial condition $x \in \mathcal{X}^r$ (i.e, such that $\Xi^r(0) = x$). Suppose that the conditions of Theorem 3.1 hold. Then, there exists a filtration $\mathcal{F}^r = (\mathcal{F}_t^r)_{t \geq 0}$ and a J -dimensional process $\widehat{\zeta}^r$ such that:*

- (i) $(\{\widehat{\xi}^r(t) + R\gamma t - \widehat{\zeta}^r(t), \mathcal{F}_t^r, t \geq 0\})$ is a martingale,
- (ii) $\limsup_{r \rightarrow \infty} \sup_{x \in \mathcal{A}^r} \mathbb{E}_x[\|\widehat{\zeta}^r\|_T^q] = 0$, for all $q, T > 0$, and
- (iii) The processes $\widehat{W}^r, \widehat{Y}^r, \widehat{\eta}^r$ and $\widehat{\zeta}^r$ are adapted to \mathcal{F}^r .

Let $\check{\xi}^r(t) = \widehat{\xi}^r(t) + \gamma t - \widehat{\zeta}^r(t)$, $\check{X}^r = R\check{\xi}^r$ and define

$$\check{W}^r(t) = \widehat{W}^r(0) + \check{X}^r(t) - R\gamma t + R\widehat{Y}^r(t).$$

Then \check{W}^r is, by Lemma 5.8, a semi-martingale with respect to \mathcal{F}^r . By Ito's formula we have, for all $t \geq 0$, that

$$\begin{aligned} \Psi(\check{W}^r(t)) &= \Psi(\check{W}^r(0)) + \int_0^t D\Psi(\check{W}^r(s-)) \cdot (-R\gamma) ds + \sum_{i=1}^J \int_0^t D\Psi(\check{W}^r(s-)) \cdot R^i d\widehat{Y}_i^r(s) \quad (63) \\ &+ \int_0^t D\Psi(\check{W}^r(s-)) \cdot d\check{X}^r(s) + \vartheta^r(t), \end{aligned}$$

where

$$\begin{aligned} \vartheta^r(t) &= \sum_{s \leq t} \left(\Psi(\check{W}^r(s)) - \Psi(\check{W}^r(s-)) - \sum_{i=1}^J \partial_i \Psi(\check{W}^r(s-)) \Delta \check{W}_i^r(s) \right. \\ &\quad \left. - \frac{1}{2} \sum_{i,l} \partial_i \partial_l \Psi(\check{W}^r(s-)) \Delta \check{W}_i^r(s) \Delta \check{W}_l^r(s) \right). \end{aligned}$$

Next, let κ be a constant such that $\|D^2\Psi(w)\| \leq \epsilon$ for all w with $\|w\| \geq \kappa$ (see property (P3) of $\Psi(\cdot)$). Given $T > 0$, define the stopping time $\tau_T^r = \inf\{t \geq 0 : \|\check{W}^r(t)\| \leq \kappa\} \wedge T$.

Lemma 5.9 *Suppose that the conditions of Theorem 3.1 hold and fix $\epsilon > 0$. Then, There exists an absolute constant ε , not depending on ϵ , such that, for all $r \in \mathbb{N}$,*

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}[\|\vartheta^r\|_{\tau_T^r}] \leq \varepsilon \epsilon.$$

Lemma 5.10 *Suppose that the conditions of Theorem 3.1 hold and fix $\epsilon > 0$. Then, there exists an absolute constant ε , not depending on ϵ , such that, for all sufficiently large $r \in \mathbb{N}$,*

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\sum_{i=1}^J \int_0^{t \wedge \tau_T^r} \left| \left(D\Psi(\check{W}^r(s-)) - D\Psi(\widehat{W}^r(s-)) \right) \cdot R^i \right| d\widehat{Y}_i^r(s) \right] \leq \varepsilon \epsilon,$$

and

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\sum_{i=1}^J \int_0^{t \wedge \tau_T^r} \left| \left(D\Psi(\check{W}^r(s-)) - D\Psi(\widehat{W}^r(s-)) \right) \right| ds \right] \leq \varepsilon \epsilon.$$

Lemma 5.11 *Suppose that the conditions of Theorem 3.1 hold and fix $\epsilon, t > 0$ and $q \in \mathbb{N}$. Then, there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,*

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\left(\left| \Psi(\widehat{W}^r(t)) - \Psi(w) \right| \right)^q \right] \leq \varepsilon. \quad (64)$$

Proof of Proposition 5.2: We prove first that $\Phi(\cdot) = b + \Psi(\cdot)$ is a constrained Lyapunov function of order $q = 1$ for the process $\widehat{\Xi}^r$. We then apply Proposition 5.3 to extend the conclusion to $q > 1$.

Observe that \widehat{Y}^r is Lipschitz continuous, $\Psi(\cdot)$ is continuous and the number of jumps of \widehat{W}^r on $[0, T]$ (corresponding to jumps of the underlying renewal processes) is almost surely finite. Thus, almost surely,

$$\int_0^{t \wedge \tau_T^r} D\Psi(\widehat{W}^r(s-)) \cdot R^i d\widehat{Y}_i^r(s) = \int_0^{t \wedge \tau_T^r} D\Psi(\widehat{W}^r(s)) \cdot R^i d\widehat{Y}_i^r(s).$$

Plugging Lemmas 5.9 and 5.10 into (63) and recalling that \check{X}^r is a martingale, we have for $x \in \mathcal{A}_\epsilon^r$ that

$$\begin{aligned} &\mathbb{E}_x[\Psi(\check{W}^r(t \wedge \tau_T^r))] - \mathbb{E}_x[\Psi(\check{W}^r(0))] \quad (65) \\ &\leq \mathbb{E}_x \left[\int_0^{t \wedge \tau_T^r} D\Psi(\widehat{W}^r(s-)) \cdot (-R\gamma) ds \right] + \sum_{i=1}^J \mathbb{E}_x \left[\int_0^{t \wedge \tau_T^r} D\Psi(\widehat{W}^r(s)) \cdot R^i d\widehat{Y}_i^r(s) \right] + c_2 \epsilon \\ &\leq -c_3 \mathbb{E}_x[\tau_T^r \wedge t] + c_2 \epsilon, \end{aligned}$$

where c_2, c_3 are absolute constants that do not depend on ϵ . Here, the last inequality follows from property (P4) of the function $\Psi(\cdot)$ recalling (12) that \widehat{Y}_i^r can increase only at times t in which $\widehat{W}_i^r(t) = 0$.

Fix $t_0 > 8c_2\epsilon/c_3$. For sufficiently large κ and all $r \in \mathbb{N}$, it holds that

$$\sup_{x \in \mathcal{A}_\epsilon^r, \Psi(w) \geq 2\kappa} |\mathbb{E}_x[t_0 \wedge \tau_T^r] - t_0| \leq t_0/8, \quad (66)$$

and

$$\sup_{x \in \mathcal{A}_\epsilon^r, \Psi(w) \geq 2\kappa} |\mathbb{E}_x[\Psi(\check{W}^r(t_0 \wedge \tau_T^r))] - \mathbb{E}_x[\Psi(\check{W}^r(t_0))]| \leq c_3 t_0/8. \quad (67)$$

The proofs of (66) and (67) appear in §F of the appendix. Combining (65)-(67) we obtain

$$\sup_{x \in \mathcal{A}_\epsilon^r, \Psi(w) \geq 2\kappa} \frac{\mathbb{E}_x[\Psi(\check{W}^r(t_0))] - \Psi(w)}{t_0} \leq -c_3 t_0/2.$$

By Lemma 5.8 $\|\check{W}^r - \widehat{W}^r\|_T \leq \|\zeta^r\|_T + \|R\hat{\eta}^r\|_T$ where $\limsup_{r \rightarrow \infty} \sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x[\|\zeta^r\|_T^q] = 0$. From Theorem 5.4 it then follows that, for sufficiently large $r \in \mathbb{N}$, $\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x[\|\check{W}^r - \widehat{W}^r\|_T^q] \leq c_4\epsilon$ where c_4 does not depend on ϵ . Re-choosing $\epsilon \leq c_3 t_0/(4c_4)$ we conclude that, for all sufficiently large $r \in \mathbb{N}$,

$$\sup_{x \in \mathcal{A}_\epsilon^r, \Psi(w) \geq 2\kappa} \frac{\mathbb{E}_x[\Psi(\widehat{W}^r(t_0))] - \Psi(w)}{t_0} \leq -c_3 t_0/4,$$

and, in particular, that $\Psi(\cdot)$ is a constrained Lyapunov function of order $q = 1$ for all such r , with drift-size parameter $c_3 t_0/4$ drift-time parameter t_0 , exception parameter 2κ , and inclusion set \mathcal{A}_ϵ^r . The function $\Phi(\cdot)$ (see (41)) is then itself a constrained Lyapunov function of order $q = 1$ with these parameters and this inclusion set.

To show that $\Phi(\cdot)$ is, in fact, a constrained Lyapunov function of order $q > 1$ we next verify that the conditions of Proposition 5.3 hold. By Lemma 5.11,

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[(\Psi(\widehat{W}^r(t)) - \Psi(w))^2 (b + \Psi(w) + |\Psi(\widehat{W}^r(t)) - \Psi(w)|)^{q-2} \right] \leq c_6 + c_7 (b + \Psi(w))^{q-2}.$$

Recalling the definitions (41) and (62) we then have $L_q^{\widehat{\epsilon}^r}(t, \mathcal{A}_\epsilon^r) \leq c_8$. It is similarly argued that $\phi_q^{\widehat{\epsilon}^r}(t^0, \mathcal{A}_\epsilon^r) \leq c_9$. Thus,

$$\max\{\phi_q^{\widehat{\epsilon}^r}(t_0, \mathcal{A}_\epsilon^r), L_q^{\widehat{\epsilon}^r}(t_0, \mathcal{A}_\epsilon^r)\} \leq c_{10}, \quad (68)$$

for all $r \in \mathbb{N}$. From Proposition 5.3 it then follows that $\Phi(\cdot)$ is a constrained Lyapunov function of order $q > 1$ with the appropriate parameters. Equation (68) establishes also the bound (42) and concludes the proof of the proposition. \blacksquare

6. Concluding remarks

6.1 Summary The analysis of the limit-interchange problem for multiclass queue-ratio networks underscores interesting connections between fluid-models, stability, state-space collapse and tightness of the stationary distributions. As in the analysis of stability and state-space collapse, fluid models are shown to play a crucial role also for the limit-interchange problem. The complexity of multiclass queuing networks, particularly the lack of a continuous mapping from model primitives to workload, motivates the introduction of two new concepts: *constrained Lyapunov functions*, and *truncated fluid models*.

The analysis of queue-ratio networks suggests a road-map for establishing the validity of steady-state heavy-traffic approximations that consists of three main steps:

- (i) verifying that the queueing network is stable;
- (ii) verifying that the diffusion limit is stable and that one can identify a Lyapunov function for this diffusion limit; and
- (ii) verifying that there is a linear attraction of a suitably defined fluid model to the invariant manifold.

Provided that these pre-requisites are satisfied it may be possible to establish limit-interchange by making use of the Lyapunov function for the diffusion limit as a constrained Lyapunov function for the sequence of queueing networks.

We conjecture that this approach is executable for several well studied settings including (i) maximum pressure policies (see e.g. [1]), (ii) max-weight type controls in parallel-server settings of which a well known instance is the Generalized $c\mu$ ($Gc\mu$) control studied in [28] or (iii) bandwidth sharing setting as in [25]. These seem to share with queue-ratio disciplines desirable characteristics that will facilitate the deployment of our framework towards proving limit-interchange. These desirable properties are a “fast” attraction to the invariant manifold and a simplifying structure of the fluid model that allows for an analogue of Lemma 2.1 relating the attraction of the fluid model to the invariant manifold to that of its truncated counterpart.

At the same time, there are disciplines for which it is *unlikely* that our framework may be applied, these are (i) disciplines that use customer service time information in making allocation decisions, and (ii) disciplines that exhibit slow convergence to the invariant manifold.

Service time information: In our construction of the sample paths service times are realized only when service on a job commences. This precludes disciplines that use a priori service time information, such as Shortest Processing Time First (SPTF) or disciplines that use workload information. This restriction is driven by our use of Lyapunov functions techniques in our proofs.

“Slow” convergence to the invariant manifold A crucial requirement in our results is that of linear attraction to the invariant manifold. It is *unlikely* that our analysis can be extended to disciplines which exhibit sub-linear convergence. Important examples of such disciplines are FIFO and HLPPS. The main obstacle in applying our approach to FIFO networks with feedback is that, as shown by Bramson [4], the rate of convergence of fluid solutions to the invariant manifold depends on the initial condition (beyond its mere distance from the invariant manifold). In our approach it is necessary to show that if the network is initialized in a small neighborhood of the invariant manifold, it will not “drift away”.

As the discussion after Theorem 4.1 suggests, it may be possible to achieve the diffusion limits of FIFO, HLPPS and other disciplines, using a queue-ratio discipline with appropriately chosen lifting matrix Δ . As we have shown in this paper, queue-ratio disciplines provide an added technical benefit, insofar as, conditions for purposes of limit interchange all assumption can be made with respect to well-studied fluid models.

Appendix The appendix includes the proofs of all results that were stated in the main part of the manuscript without proof. Section A includes some preliminary probability bounds. In §B we prove some Lyapunov-function results including Propositions 5.1 and 5.3. Section C is dedicated to the proof of Theorem 5.6. In sections D and E we prove Theorems 5.4 and 5.3, respectively. Finally, §F includes proofs of various auxiliary lemmas.

In all of §C-E, our proofs rely on the state-space-collapse framework developed by Bramson in [5]. Many of the arguments in [5] can be imported without change to our setting and, in those instances, we will be succinct while making the necessary references to [5]. Some familiarity with that work is thus assumed.

Throughout the appendix, rather than assuming that (4) holds for all p , we keep explicit the dependence of the various bounds on the number of moments that exist for the interarrival and service times. Throughout p will be an integer for which (4) holds. Finally, recall that we use the term *absolute constant* to denote a finite and strictly positive constant that does not depend on the heavy-traffic index r (but that may depend on various other parameters). We use c_0, c_1, c_2, \dots to denote such constants.

Appendix A. Some bounds on the primitives Let $\Pi^r = \text{diag}(\alpha_1^r c_{a,1}^2, \dots, \alpha_K^r c_{a,K}^2)$. The matrices Σ and Υ are defined as in §2.1. Recall that the service-time distributions and the routing matrix P do not scale with r , so that we may write $S^r(t) = S(rt)$ and $\Phi^r(\lfloor rt \rfloor) = \Phi(\lfloor rt \rfloor)$ for all $r \in \mathbb{N}$ and $t \geq 0$, and for fixed processes S and Φ as defined in §2.1.

Lemma A.1 (strong approximations) Fix a function $a(\cdot)$ with $a(\theta)/\theta \rightarrow 0$ as $\theta \rightarrow \infty$. Then, there exist $K + 2$ mutually independent standard K -dimensional Brownian motions $B^k = (B^k(t) : t \geq 0)$, $k = 0, \dots, K + 1$, such that, for all $x \in \mathcal{X}^r$, and $y \in [c_1 \log(rT), a(rT)]$

$$\mathbb{P}_x \left\{ \sup_{rT \wedge \mathcal{R}_k^{a,r}(0) \leq t \leq rT} \left| E_k^r(t - \mathcal{R}_k^{a,r}(0)) - \alpha_k^r(t - \mathcal{R}_k^{a,r}(0)) - \sqrt{\Pi_k^r} B_k^0(t - \mathcal{R}_k^{a,r}(0)) \right| > y \right\} \leq \frac{c_k^a rT}{y^p},$$

$$\mathbb{P}_x \left\{ \sup_{rT \wedge \mathcal{R}_k^{v,r}(0) \leq t \leq rT} \left| S_k(t - \mathcal{R}_k^{v,r}(0)) - \mu_k(t - \mathcal{R}_k^{v,r}(0)) - \sqrt{\Sigma_k} B_k^{K+1}(t - \mathcal{R}_k^{v,r}(0)) \right| > y \right\} \leq \frac{c_k^s rT}{y^p},$$

$$\mathbb{P}_x \left\{ \sup_{0 \leq t \leq rT} \left| \varphi_l^k(\lfloor t \rfloor) - p_{kl}t - \sqrt{\Upsilon_{kl}} B_l^k(t) \right| > y \right\} \leq \frac{c_{k,l} rT}{y^p},$$

for $k, l = 1, \dots, K$. Here c_k^a , c_k^s and $c_{k,l}$, $k, l = 1, \dots, K$ are absolute constants.

Lemma A.1 is a direct corollary of Theorem 2.2.7 in Csörgo and Horváth [13]. We use it to prove the following proposition. Below \mathcal{B}_ϵ^r is as in (25) and the scaled processes \widehat{E}^r , \widehat{S}^r and $\widehat{\varphi}^r$ are as defined in §4.

Proposition A.1 Fix $\epsilon > 0$ and $q < p - 1$. Then, there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{E}_x \left[\left(\|\widehat{E}^r\|_T \vee \|\widehat{\varphi}^r\|_T \vee \|\widehat{S}^r\|_T \right)^q \right] \leq \varepsilon T^{\frac{q}{2}}.$$

Proof: We prove the bound for \widehat{E}^r . The bounds for the other processes are proved similarly. Fix $\eta \in (0, \epsilon)$ with $\eta \leq (p - 1 - q)/p$. We will show that, for $y \in [c_l \log(rT)/\sqrt{rT}, (rT)^{1-\eta-1/2}]$,

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{P}_x \left\{ \sup_{0 \leq t \leq rT} \left| E_k^r(t) - \alpha_k^r t - \sqrt{\Pi_k^r} B_k^0(t) \right| > y\sqrt{rT} \right\} \leq \frac{c_3 rT}{y^p (rT)^{p/2}}, \quad (69)$$

where $\{B_k^0, k = 1, 2, \dots, K\}$ are as in Lemma A.1. To that end, we re-write, for $t \geq \mathcal{R}_k^{a,r}(0)$,

$$E_k^r(t) - \alpha_k^r t - \sqrt{\Pi_k^r} B_k^0(t) = E_k^r(t - \mathcal{R}_k^{a,r}(0)) - \alpha_k^r(t - \mathcal{R}_k^{a,r}(0)) - \sqrt{\Pi_k^r} B_k^0(t - \mathcal{R}_k^{a,r}(0)) + 1 - \alpha_k^r \mathcal{R}_k^{a,r}(0)$$

The function $a(\theta) = \theta^{1-\eta}$ satisfies the condition of Lemma A.1 so that, since $|1 - \alpha_k^r \mathcal{R}_k^{a,r}(0)| \leq 1 + 2\alpha_k r^{-\eta}$ for all $x \in \mathcal{B}_\epsilon^r$ and $y\sqrt{rT} \geq \log(rT)$ for all $y \in [c_l \log(rT)/\sqrt{rT}, (rT)^{1-\eta-1/2}]$, we have that

$$\begin{aligned} & \sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{P}_x \left\{ \sup_{0 \leq t \leq rT} \left| E_k^r(t) - \alpha_k^r t - \sqrt{\Pi_k^r} B_k^0(t) \right| > y\sqrt{rT} \right\} \\ & \leq \sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{P}_x \left\{ \sup_{rT \wedge \mathcal{R}_k^{a,r}(0) \leq t \leq rT} \left| E_k^r(t - \mathcal{R}_k^{a,r}(0)) - \alpha_k^r(t - \mathcal{R}_k^{a,r}(0)) - \sqrt{\Pi_k^r} B_k^0(t - \mathcal{R}_k^{a,r}(0)) \right| > \frac{1}{2} y\sqrt{rT} \right\} \\ & \leq \frac{c_3 rT}{y^p (rT)^{p/2}}. \end{aligned}$$

This establishes (69). Next, let

$$\mathcal{L}_\eta^r = \left\{ \omega \in \Omega : \sup_{0 \leq t \leq rT} \left| E_k^r(t) - \alpha_k^r t - \sqrt{\Pi_k^r} B_k^0(t) \right| \leq (rT)^{1-\eta} \right\}.$$

Since $q < p - 1$ we have, integrating (69), that

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{E}_x \left[\left(\sup_{0 \leq t \leq rT} \frac{1}{\sqrt{r}} \left| E_k^r(t) - \alpha_k^r t - \sqrt{\Pi_k^r} B_k^0(t) \right| \right)^q \mathbb{1}\{\mathcal{L}_\eta^r\} \right] \leq c_4 \frac{T^{\frac{q}{2}}}{(rT)^{p/2-1}}. \quad (70)$$

Evidently,

$$\begin{aligned} \mathbb{E}_x \left[\left(\sup_{0 \leq t \leq rT} \left| E_k^r(t) - \alpha_k^r t - \sqrt{\Pi_k^r} B_k^0(t) \right| \right)^q \mathbb{1}\{(\mathcal{L}_\eta^r)^c\} \right] &\leq \\ \mathbb{E}_x [(E_k(rT))^q \mathbb{1}\{(\mathcal{L}^r)^c\}] &\leq \sqrt{\mathbb{E}_x [(E_k(rT))^{2q}]} \sqrt{\mathbb{P}_x\{(\mathcal{L}_\eta^r)^c\}}, \end{aligned} \quad (71)$$

where the last step follows from Hölder's inequality. By known bounds for renewal processes (see e.g. Lemma 5.2 in [16]) we have that

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{E}_x [(E_k^r(t))^l] \leq c_5 t^l + c_5. \quad (72)$$

By (69) $\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{P}_x\{(\mathcal{L}_\eta^r)^c\} \leq c_6 r^{-(p(1-\eta)-1)}$. Plugging this and (72) into (71) and recalling that $\eta < (p - q - 1)/p$ we have

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{E}_x \left[\left(\frac{1}{\sqrt{r}} \sup_{0 \leq t \leq rT} \left| E_k(t) - \alpha_k^r t - \sqrt{\Pi_k^r} B_k^0(t) \right| \right)^q \mathbb{1}\{(\mathcal{L}_\eta^r)^c\} \right] \leq c_7 T^{\frac{q}{2}} (rT)^{\frac{q}{2} - \frac{1}{2}(p(1-\eta)-1)} \leq c_7 T^{\frac{q}{2}}.$$

Consequently, the Brownian motion \tilde{B}_k^0 defined by $\tilde{B}_k^0(t) = B_k^0(rt)/\sqrt{r}$, satisfies

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{E}_x \left[\left(\sup_{0 \leq t \leq T} \left| \hat{E}_k^r(t) - \sqrt{\Pi_k^r} \tilde{B}_k^0(t) \right| \right)^q \right] \leq c_8 T^{\frac{q}{2}}. \quad (73)$$

Finally, by basic properties of Brownian motion $\mathbb{E} \left[\|\tilde{B}_k^0\|_T^q \right] \leq c_9 T^{\frac{q}{2}}$ and we have, as required, that

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{E}_x \left[\|\hat{E}_k^r\|_T^q \right] \leq c_{10} T^{\frac{q}{2}}.$$

This concludes the proof. ■

Appendix B. More Lyapunov-function tools As before, when making a reference to a Lyapunov function this is always with respect to an underlying Markov process $\hat{\Xi} = (\hat{\Xi}(t), t \geq 0)$ on a locally compact separable metric state space $\hat{\mathcal{X}}$. We start with some results for regular (i.e. un-constrained) Lyapunov functions.

Definition 5 (Lyapunov function of order q) A function $\Phi : \hat{\mathcal{X}} \rightarrow \mathbb{R}_+$ is said to be a Lyapunov function of order $q \geq 1$ with drift-size parameter $-\delta < 0$, drift-time parameter $t_0 > 0$ and exception parameter κ , if

$$\sup_{x \in \hat{\mathcal{X}}: \Phi(x) > \kappa} \frac{\mathbb{E}_x[\Phi^q(\hat{\Xi}(t_0)) - \Phi^q(x)]}{\Phi^{q-1}(x)} \leq -\delta. \quad (74)$$

Given a Lyapunov function $\Phi : \hat{\mathcal{X}} \rightarrow \mathbb{R}_+$ and $t \geq 0$, let

$$\phi_q^{\hat{\Xi}}(t) = \sup_{x \in \hat{\mathcal{X}}} \left\{ \Phi^{-(q-1)}(x) \mathbb{E}_x[(\Phi^q(\hat{\Xi}(t)) - \Phi^q(x))^+] \right\}, \quad (75)$$

and

$$L_q^{\hat{\Xi}}(t) = \sup_{x \in \hat{\mathcal{X}}} \left\{ \Phi^{-(q-2)}(x) \mathbb{E}_x \left[\Phi(\hat{\Xi}(t)) - \Phi(x) \right]^2 (\Phi(x) + |\Phi(\hat{\Xi}(t)) - \Phi(x)|)^{q-2} \right\}, \quad (76)$$

where $\phi_q^{\hat{\Xi}}(t)$ and $L_q^{\hat{\Xi}}(t)$ may be infinite.

Proposition B.1 *Fix $q > 1$. Suppose that Φ is a Lyapunov function of order 1 with parameters δ, t_0 and κ . Then, provided that $L_q^{\hat{\Xi}}(t_0)$ is finite, $\Phi(\cdot)$ is a Lyapunov function of order q with drift-size parameter $-\delta q/2$, drift-time parameter t_0 , and exception parameter $\max\{\kappa, L_q^{\hat{\Xi}}(t_0)(q-1)/\delta\}$.*

Proof: Using second order Taylor's expansion of the function $g(y) = y^q$ around $\Phi(x)$ we obtain for every $x \in \widehat{\mathcal{X}}$ such that $\Phi(x) > \kappa$,

$$\begin{aligned} \mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))] - \Phi^q(x) &= q\Phi^{q-1}(x)\mathbb{E}_x[\Phi(\widehat{\Xi}(t_0)) - \Phi(x)] \\ &\quad + \frac{q(q-1)}{2}\mathbb{E}_x[(\Phi(x) + Z(\Phi(\widehat{\Xi}(t_0)) - \Phi(x)))^{q-2}(\Phi(\widehat{\Xi}(t_0)) - \Phi(x))^2] \\ &\leq -\gamma q\Phi^{q-1}(x) + \frac{q(q-1)}{2}\mathbb{E}_x[(\Phi(x) + |\Phi(\widehat{\Xi}(t_0)) - \Phi(x)|)^{q-2}(\Phi(\widehat{\Xi}(t_0)) - \Phi(x))^2] \\ &\leq -\gamma q\Phi^{q-1}(x) + \frac{q(q-1)}{2}L_q^{\widehat{\Xi}}(t_0)\Phi^{q-2}(x) \end{aligned}$$

where Z is a random variable whose support is contained in the interval $[0, 1]$. When $\Phi(x) > L_q^{\widehat{\Xi}}(t_0)(q-1)/\delta$, we obtain that

$$\mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))] - \Phi^q(x) \leq -\frac{\delta q}{2}\Phi^{q-1}(x).$$

This concludes the proof. ■

Proposition B.2 *Suppose that the Markov process $\widehat{\Xi}$ possesses a stationary distribution π and that Φ is a Lyapunov function of order q with parameters δ , t_0 and κ . Then,*

$$\mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \right] \leq \kappa^{q-1} \left(1 + \frac{\phi_q(t_0)}{\delta} \right). \quad (77)$$

Proof: Our proof draws on the proof of Theorem 5 in [20]. Since $\Phi(\cdot)$ is assumed to be a Lyapunov function of order q , we have that

$$\mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))] - \Phi^q(x) \leq -\delta\Phi^{q-1}(x)\mathbb{1}\{\Phi(x) > \kappa\} + \phi_q(t_0)\kappa^{q-1}\mathbb{1}\{\Phi(x) \leq \kappa\}, \quad (78)$$

for any $x \in \widehat{\mathcal{X}}$. Fix $m \in \mathbb{N}$ and put $\Phi_m(x) = \Phi(x) \wedge m$. Since π is a stationary distribution for $\widehat{\Xi}$, we then have that

$$\int (\Phi_m^q(x) - \mathbb{E}_x[\Phi_m^q(\widehat{\Xi}(t_0))])\pi(dx) = 0.$$

By (78)

$$\Phi_m^q(x) - \mathbb{E}_x[\Phi_m^q(\widehat{\Xi}(t_0))] \geq -\phi_q(t_0)\kappa^{q-1},$$

whenever $\Phi^q(x) \leq m$. Also, if $\Phi^q(x) > m$ we have that

$$\Phi_m^q(x) - \mathbb{E}_x[\Phi_m^q(\widehat{\Xi}(t_0))] \geq m - \mathbb{E}_x[(\Phi_m^q(\widehat{\Xi}(t_0)))] \geq 0,$$

so that $\Phi_m^q(x) - \mathbb{E}_x[\Phi_m^q(\widehat{\Xi}(t_0))]$ is bounded from below by $-\phi_q(t_0)\kappa^{q-1}$ for all $x \in \widehat{\mathcal{X}}$. We apply Fatou's lemma to conclude that

$$\begin{aligned} \int (\Phi^q(x) - \mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))])\pi(dx) &\leq \liminf_{m \rightarrow \infty} \int (\Phi_m^q(x) - \mathbb{E}_x[\Phi_m^q(\widehat{\Xi}(t_0))])\pi(dx) \\ &= 0. \end{aligned} \quad (79)$$

Plugging (78) into (79) we then have that

$$\delta\mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0))\mathbb{1}\{\Phi(\widehat{\Xi}(0)) > \kappa\} \right] \leq \kappa^{q-1}\phi_q(t_0),$$

and, in particular, that

$$\mathbb{E}_\pi \left[\Phi^{q-1}(\widehat{\Xi}(0)) \right] \leq \kappa^{q-1} \left(1 + \frac{\phi_q(t_0)}{\delta} \right),$$

as required. ■

Proof of Proposition 5.1: By assumption, $\mathbb{E}_\pi[\Phi^q(\widehat{\Xi}(0))] < \infty$, so that

$$\int (\Phi^q(x) - \mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))])\pi(dx) = 0.$$

Consequently,

$$\int_{\mathcal{A}} (\Phi^q(x) - \mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))])\pi(dx) + \int_{\mathcal{A}^c} (\Phi^q(x) - \mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))])\pi(dx) = 0. \quad (80)$$

As $\Phi(\cdot)$ is a constrained Lyapunov function we have, by condition (c) of the proposition, that

$$\Phi^q(x) - \mathbb{E}_x[\Phi^q(\widehat{\Xi}(t_0))] \geq \delta\Phi^{q-1}(x) - \kappa^{q-1}\phi_q(t_0, \mathcal{A}),$$

for all $x \in \mathcal{A}$. The assertion in (38) now follows from (80) through simple algebraic manipulations. ■

Proof of Proposition 5.3: The proof is identical to that of Proposition B.1. ■

Appendix C. Crude steady-state bounds Given an initial condition $x \in \mathcal{X}^r$ (i.e. such that $\Xi^r(0) = x$) define, for $t \geq 0$,

$$\begin{aligned} Z^{r,x}(t) &= Z^r(\|x\|\sqrt{rt}), \quad W^{r,x}(t) = CMZ^r(\|x\|\sqrt{rt}), \quad Y^{r,x}(t) = Y^r(\|x\|\sqrt{rt}), \\ \mathcal{R}^{a,r,x}(t) &= \mathcal{R}^{a,r}(\|x\|\sqrt{rt}), \quad \mathcal{R}^{v,r,x}(t) = \mathcal{R}^{v,r}(\|x\|\sqrt{rt}), \end{aligned}$$

and let

$$\Xi^{r,x} = (Z^{r,x}, \mathcal{R}^{a,r,x}, \mathcal{R}^{v,r,x}).$$

The dependence of $\Xi^{r,x}$ on both the heavy-traffic r and the initial condition x is central to our proof of Theorem 5.6. The rest of this section is structured as follows: to establish fluid limits that hold uniformly in r and x , we first prove a suitable SSC result for the doubly-indexed process $\Xi^{r,x}/(\|x\| \vee 1)$; see §C.1. We proceed in §C.2 to prove that the process $\widehat{W}^{r,x} = W^{r,x}/(\|x\| \vee 1)$ is close, in an appropriate sense, to a Skorohod problem (SP). The assumed attraction of the (SP) to the origin (see Definition 4 and Assumption 2) is invoked in Corollary C.5 to conclude the following “downward drift”:

$$\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \mathbb{E}_x [\|\Xi^{r,x}(t_0)\|^{q+1}] \leq \frac{1}{2} \|x\|^{q+1}, \quad (81)$$

for some $\delta, t_0 > 0$, all sufficiently large r and any integer $q < p$. A downward drift condition similar to (81) is the starting point of the analysis in [16]; see the bottom of page 1898 there. Once (81) has been established, the bound in Theorem 5.6 will follow from [16]. This is argued in §C.3.

C.1 Uniform SSC results Given $r \in \mathbb{N}$ and an initial condition $x \in \mathcal{X}^r$, let

$$\widehat{\Xi}^{r,x} = \frac{\Xi^{r,x}}{\|x\| \vee 1}.$$

Namely, $\widehat{\Xi}^{r,x}$ is obtained from $\Xi^{r,x}$ by further dividing each of its components by $\|x\| \vee 1$ and we write

$$\widehat{\Xi}^{r,x} = (\widehat{Z}^{r,x}, \widehat{\mathcal{R}}^{a,r,x}, \widehat{\mathcal{R}}^{v,r,x}).$$

By definition, $\widehat{W}^{r,x} = CM\widehat{Z}^{r,x}$ and $\widehat{\epsilon}^{r,x} = \widehat{Z}^{r,x} - \Delta\widehat{W}^{r,x}$.

We say that the fluid model (16)-(22) is *attracted to the invariant manifold* if, given $\epsilon > 0$, there exists $t_0(\epsilon)$ such that, $\|\bar{\epsilon}(t)\| \leq \epsilon$, for all $t \geq t_0(\epsilon)$, and any solution $\bar{X} = (\bar{W}, \bar{Z}, \bar{\epsilon}, \bar{T})$ to (16)-(22). The linear attraction assumed in Theorem 5.1, implies, in particular, this weaker form of attraction.

Theorem C.1 *Fix $\delta, \epsilon, T > 0$. Assume that the fluid model (16)-(22) is attracted to the invariant manifold in finite time. Then, there exist absolute constants ε, s such that, for all $r \in \mathbb{N}$,*

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{\|\widehat{\epsilon}^{r,x}\|_{s/\sqrt{r}, T}}{\|\widehat{W}^{r,x}\|_T \vee 1} > \epsilon \right\} \leq \frac{\varepsilon}{r^{p-2-\epsilon p}}.$$

The next corollary strengthens the multiplicative SSC in Theorem C.1. This is analogous to Proposition 8.1 in [38]. The proof appears in §F

Corollary C.2 Fix $\delta, \epsilon, T > 0$. Assume that the fluid model (16)-(22) is attracted to the invariant manifold in finite time. Then, there exist absolute constants ε, s such that, for all $r \in \mathbb{N}$,

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|\hat{c}^{r,x}\|_{s/\sqrt{r}, T} > \epsilon \right\} \leq \frac{\varepsilon}{r^{p-2-\epsilon p}}.$$

The rest of this section is dedicated to the proof of Theorem C.1. We first construct the so-called hydrodynamically-scaled processes. Given $r, m \in \mathbb{N}$, and $x \in \mathcal{X}^r$ let

$$y^{r,m,x} = \|Z^{r,x}(m\|x\|)\| \vee \|W^{r,x}(m\|x\|)\| \vee \|x\| \vee 1.$$

For a process $w \in \mathcal{D}^d[0, \infty)$ and given $y, \iota \geq 0$ and $m \in \mathbb{N}$, let

$$\mathfrak{D}^{y,\iota}[w](t) = \frac{1}{y} \left(w(m\iota + yt) - w(m\iota) \right). \quad (82)$$

Define

$$E^{r,m,x} = \mathfrak{D}^{y^{r,m,x}, \|x\|}[E^r], \quad D^{r,m,x} = \mathfrak{D}^{y^{r,m,x}, \|x\|}[D^r],$$

and

$$T^{r,m,x} = \mathfrak{D}^{y^{r,m,x}, \|x\|}[T^r].$$

As in [5], the process $\Phi^{r,m,x}$ is constructed by re-starting the clock at time $m\|x\|$ and scaling time and space appropriately. Namely, we create $\lceil \sqrt{r}T \rceil$ independent copies of Φ^r , with the m^{th} copy given by $\Phi^{r,m}$ so that

$$\Phi^{r,m,x}(D^{r,m,x}(t)) = \frac{1}{y^{r,m,x}} \left(\Phi^{r,m}(D^r(m\|x\| + y^{r,m,x}t) - D^r(m\|x\|)) \right).$$

Finally, we set

$$Z^{r,m,x}(t) = \frac{Z^r(m\|x\| + y^{r,m,x}t)}{y^{r,m,x}}, \quad \text{and} \quad W^{r,m,x}(t) = \frac{W^r(m\|x\| + y^{r,m,x}t)}{y^{r,m,x}},$$

and put

$$\mathfrak{X}^{r,m,x} = (Z^{r,m,x}, W^{r,m,x}, E^{r,m,x}, D^{r,m,x}, T^{r,m,x}, \Phi^{r,m,x}).$$

For fixed m , this hydrodynamic scaling is similar to the fluid scaling used in [14] and [16] where both space and time are scaled by the “size” $\|x\|$ of the initial condition. Here, we apply that fluid scaling simultaneously to multiple time intervals. In what follows the reader will note that we consider only $m \geq 1$. This is driven by the fact that the residual service and inter-arrival times at time 0 may be large. It holds, however, that $\|R^{a,r}(0)\| + \|R^{a,r}(0)\| \leq \|x\|$ where $x = \Xi^r(0)$ which guarantees that the system “restarts” by $m = 1$.

Let

$$N = 2(2K + K^2) \left(\max_{k \in \mathcal{K}} \{\mu_k\} + \max_{k \in \mathcal{K}^a} \{\alpha_k\} \right). \quad (83)$$

Fixing $L, T, \epsilon > 0$ and letting $\nu_\epsilon(r) = r^{-\epsilon}$, we define the following events on the underlying probability space:

$$\Omega_1^r = \left\{ \max_{1 \leq m < \sqrt{r}T} \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|E^{r,m,x} - a^r \cdot\|_L \leq \nu_\epsilon(r) \right\}, \quad (84)$$

$$\Omega_2^r = \left\{ \max_{1 \leq m < \sqrt{r}T} \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|D^{r,m,x} - M^{-1}T^{r,m,x}\|_L \leq \nu_\epsilon(r) \right\}, \quad (85)$$

$$\Omega_3^r = \left\{ \max_{1 \leq m < \sqrt{r}T} \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \left\| \sum_{k \in \mathcal{K}} (\varphi^k)^{r,m,x} (D_k^{r,m,x}) - \tilde{P}M^{-1}T^{r,m,x} \right\|_L \leq \nu_\epsilon(r) \right\}, \quad (86)$$

$$\Omega_4^r = \left\{ \max_{1 \leq m < \sqrt{r}T} \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \sup_{t_1, t_2 \leq L} \|\mathfrak{X}^{r,m,x}(t_2) - \mathfrak{X}^{r,m,x}(t_1)\| \leq \nu_\epsilon(r) + N|t_2 - t_1| \right\}, \quad (87)$$

where $a^r \cdot$ is the function that equals $a^r t$ at time $t \geq 0$. For each $k \in \mathcal{K}$, define

$$u_k^{r,x,T,max} = \max\{|u_k^r(i)| : U_k^r(i-1) \leq \|x\|\sqrt{r}T, i = 2, 3, \dots\},$$

and

$$v_k^{r,x,T,max} = \begin{cases} \max\{|v_k^r(i)| : V_k^r(i-1) \leq \|x\|\sqrt{r}T, i = 2, 3, \dots\}, & \text{if } \mathcal{R}_k^{v,r}(0) > 0, \\ \max\{|v_k^r(i)| : V_k^r(i-1) \leq \|x\|\sqrt{r}T, i = 1, 2, \dots\}, & \text{otherwise.} \end{cases}$$

To obtain bounds that are uniform in the initial condition, we will require some uniform probability bounds on the primitives. Lemma C.3 is the corresponding analogue of Lemma 5.1 in [5] and Proposition C.1 is the analogue of Propositions 5.1, 5.2 and Corollary 5.1 in [5]. Here,

$$\Omega^r = \bigcap_{i=1}^4 \Omega_i^r,$$

for Ω_i^r , $i = 1, \dots, 4$ as in (84)-(87).

Lemma C.3 *Fix $\epsilon, \delta, T > 0$, and $k \in \mathcal{K}$. Then, there exists an absolute constant ϵ such that, for all $r \in \mathbb{N}$,*

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \left(\frac{u_k^{r,x,T,max}}{\|x\|} \right) > \epsilon \right\} \leq \frac{\epsilon}{r^{p-3/2}},$$

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \left(\frac{v_k^{r,x,T,max}}{\|x\|} \right) > \epsilon \right\} \leq \frac{\epsilon}{r^{p-3/2}}.$$

Further, for all sufficiently large r , and all $q < p$,

$$\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \mathbb{E} \left[\left(\frac{u_k^{r,x,T,max}}{\|x\|} \right)^q \right] \leq 2\epsilon, \text{ and } \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \mathbb{E} \left[\left(\frac{v_k^{r,x,T,max}}{\|x\|} \right)^q \right] \leq 2\epsilon.$$

Proposition C.1 *There exists an absolute constant ϵ such that, for all $r \in \mathbb{N}$,*

$$\mathbb{P} \{ \Omega^r \} \geq 1 - \frac{\epsilon}{r^{p-2-\epsilon p}}.$$

The proofs of Lemma C.3 and Proposition C.1 appear in §F. We will prove Theorem C.1 by showing that, for all sufficiently large r , each of the hydrodynamically scaled processes is approximated by a cluster point. Specifically, let

$$E_r = \left\{ \mathfrak{X}^{r,m,x}(\omega, \cdot) : 1 \leq m < \sqrt{r}T, x \in \mathcal{X}^r \cap \{x : \|x\| \geq \delta r\} \right\},$$

where we added the argument ω to $\mathfrak{X}^{r,m,x}$ to make explicit the dependence on the sample point. (here and throughout, for a process $x \in \mathcal{D}^d[0, L]$ we interchangeably use x and $x(\cdot)$ to denote the process $(x(t), 0 \leq t \leq L)$). Let

$$\mathcal{E} = \{E_r, r \in \mathbb{N}\}.$$

The set E_r should not be confused with the arrival process E^r .

Let E'_L be the set of RCLL functions, y , that satisfy

$$|y(t_2) - y(t_1)| \leq N|t_2 - t_1|, \text{ for all } t_1, t_2 \in [0, L],$$

and $\|y(0)\| \leq 1$ (see pp 118-120 in [5]) where N is as in (83). By asymptotically close we mean that, for every $\epsilon > 0$, there exists $r_0 \in \mathbb{N}$ so that for every $r \geq r_0$ and $y \in E_r$, there exists $x \in E'_L$ so that $\|x - y\|_L \leq \epsilon$. Per the terminology of [5], a cluster point x of \mathcal{E} is an RCLL function such that for all $\epsilon > 0$ and $r \in \mathbb{N}$ there exists $r \geq r_0$ and $y \in E_r$ such that $\|x - y\|_L \leq \epsilon$.

The following is the analogue of Proposition 6.1 in [5] (which is stated there for FIFO and generalized in §8 there to HL disciplines) and follows directly from there.

Proposition C.2 Fix $\epsilon, \delta, L, T > 0$ and sufficiently large r . Then, for $\omega \in \Omega^r$, $1 \leq m < \sqrt{r}T$ and any $x \in \mathcal{X}^r$ with $\|x\| \geq \delta r$, we have

$$\|\mathfrak{X}^{r,m,x}(\omega, \cdot) - \tilde{\mathfrak{X}}(\cdot)\|_L \leq \epsilon,$$

for some cluster point $\tilde{\mathfrak{X}}$ of \mathcal{E} with $\tilde{\mathfrak{X}} \in E'_L$.

The next result is an analogue of Proposition 8.1 in [5]. The proof is almost identical and it is omitted.

Proposition C.3 Fix $L, T > 0$. Then, all cluster points $\tilde{\mathfrak{X}}$ of \mathcal{E} satisfy the fluid model equations (16)–(22).

The following proposition is the analogue of Proposition 6.5 in [5] and follows from Propositions C.2 and C.3. Its proof is similar to (and, in fact, simpler than) that of Proposition 6.5 in [5]. We have the additional simplification that, in our setting, $W^r = CMZ^r$ by definition, while in [5] one must relate the true workload and the queues more carefully. Thus, the proof is omitted.

Proposition C.4 Fix $\epsilon, \delta, T > 0$. Then, there exists an absolute constant s such that for all sufficiently large r and all $\omega \in \Omega^r$,

$$\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{\|\tilde{\mathfrak{C}}^{r,x}\|_{s/\sqrt{r},T}}{\|\widehat{W}^{r,x}\|_T \vee 1} \leq \epsilon.$$

Proof of Theorem C.1: Combined, Propositions C.1 and C.4 prove the theorem. ■

C.2 A uniform fluid approximation We will show that $\widehat{\Xi}^{r,x}$ is suitably approximated by a Skorohod Problem (Theorem C.4) and subsequently use that to prove a “downward drift” property (Corollary C.5). Recall that, given $\epsilon > 0$, $\nu_\epsilon(r) = r^{-\epsilon}$.

Theorem C.4 Fix $T, \delta, \epsilon > 0$. Then, there exist a sequence $\{\Omega^r, r \in \mathbb{N}\}$ of subsets of Ω and absolute constants $s, \varepsilon_0, \varepsilon_1$ such that, for all $r \in \mathbb{N}$,

$$\mathbb{P}\{\Omega^r\} \geq 1 - \frac{\varepsilon_0}{r^{p-2-\epsilon p}}, \tag{88}$$

and for all $\omega \in \Omega^r$ and all $x \in \mathcal{X}^r \cap \{\|x\| \geq \delta r\}$,

$$\|\widehat{W}^{r,x} - \widetilde{W}\|_{s/\sqrt{r},T} \leq \nu_\epsilon(r), \text{ and } \|\widehat{Y}^{r,x} - \widetilde{Y}\|_{s/\sqrt{r},T} \leq \nu_\epsilon(r),$$

where $(\widetilde{W}, \widetilde{Y})$ satisfies the SP with respect to $(\widetilde{W}(0), S, -R\gamma, R)$ and $\widetilde{W}(0) \leq \varepsilon_1$.

Corollary C.5 Fix $\delta > 0$ and $q < p - 3$. Then, there exists an absolute constant t_0 such that, for all sufficiently large r ,

$$\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \mathbb{E}_x [\|\Xi^r(\|x\|\sqrt{r}t_0)\|^{q+1}] \leq \frac{1}{2}\|x\|^{q+1}.$$

Proof: By Assumption 2, the SP with data $(\widetilde{W}(0), S, -R\gamma, R)$ is attracted to the origin in finite time. In particular, there exists t_0 such that $\|\widetilde{W}(t_0)\| \leq (1/16) \min_{k \in \mathcal{K}} m_k$ for all $t \geq t_0$ with \widetilde{W} being as in Theorem C.4. Since $\widetilde{W} = CM\widetilde{Z}$, we then have that $\|\widetilde{Z}(t_0)\| \leq 1/4$ for all $t \geq t_0$. Consequently, $\|Z^{r,x}(t)\| \leq 1/8$ for each $\omega \in \Omega^r$, $x \in \mathcal{X}^r \cap \{x : \|x\| \geq \delta r\}$ and all $t \geq t_0$. Outside of the set Ω^r , the following crude bound holds

$$\|Z^r(\|x\|\sqrt{r}t_0)\|^q \leq c_0 (\|x\|^{q+1} + \|E^r(\|x\|\sqrt{r}t_0)\|^{q+1}).$$

Using (88) and bounds as in (72)) and applying Hölder’s inequality we obtain

$$\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \mathbb{E}_x [\|Z^r(\|x\|\sqrt{r}t_0)\|^{q+1}] \leq \frac{1}{4}\|x\|^{q+1}.$$

Observe that, by definition, $\mathcal{R}_k^{u,r,x}(0) \leq \|x\|$ and $\|x\|\sqrt{r}t_0 > \|x\|$ for all sufficiently large r so that $\mathcal{R}_k^{a,r,x}(t_0) \leq u_k^{r,x,T,max}$ and similarly for $\mathcal{R}_k^{v,r,x}$. By Lemma C.3 we then have that

$$\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \mathbb{E}_x \left[(\|\mathcal{R}^{a,r,x}(t_0)\| + \|\mathcal{R}^{v,r,x}(t_0)\|)^{q+1} \right] \leq 1/4 \|x\|^{q+1},$$

thus establishing the statement of the corollary. \blacksquare

The remainder of this section is dedicated to the proof of Theorem C.4. As we pursue an approximation that is uniform over initial conditions $\{x : \|x\| \geq \delta r\}$ we cannot take limits directly. Instead, we invoke again the framework of [5] but apply it now to the single interval $[0, T]$. By (34)

$$\widehat{X}^{r,x}(t) = \widehat{W}^{r,x}(0) + R \left(\widehat{\xi}^{r,x}(t) + \widehat{\eta}^{r,x}(t) \right),$$

where

$$\widehat{\xi}^{r,x}(t) = \sqrt{r} \frac{\widehat{\xi}^r(\|x\|t/\sqrt{r})}{\|x\| \vee 1} \quad \text{and} \quad \widehat{\eta}^{r,x}(t) = \sqrt{r} \frac{\widehat{\eta}^r(\|x\|t/\sqrt{r})}{\|x\| \vee 1}, \quad (89)$$

with $\widehat{\eta}^r$ and $\widehat{\xi}^r$ as defined in (32) and (33) respectively. Define the following events:

$$\begin{aligned} \widetilde{\Omega}_1^r &= \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|\widehat{X}^{r,x}(\cdot) - \widehat{X}^{r,x}(0) - \gamma t\|_{s/\sqrt{r}, T} \leq \nu_\epsilon(r) \right\}, \\ \widetilde{\Omega}_2^r &= \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \sup_{s/\sqrt{r} \leq t_1, t_2 \leq T} \|\widehat{W}^{r,x}(t_2) - \widehat{W}^{r,x}(t_1)\| \leq \nu_\epsilon(r) + N|t_2 - t_1| \right\}, \\ \widetilde{\Omega}_3^r &= \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \sup_{s/\sqrt{r} \leq t_1, t_2 \leq T} \|\widehat{Y}^{r,x}(t_2) - \widehat{Y}^{r,x}(t_1)\| \leq \nu_\epsilon(r) + N|t_2 - t_1| \right\}, \\ \widetilde{\Omega}_4^r &= \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|\widehat{W}^{r,x}\|_{s/\sqrt{r}} \leq \widetilde{\epsilon} \right\} \end{aligned}$$

where $s > 0$ is as in Corollary C.2 and $\widetilde{\epsilon}$ is an absolute constant. Let $\widetilde{\Omega}^r = \bigcap_{i=1}^4 \widetilde{\Omega}_i^r$.

Lemma C.6 *Fix $\delta, T, \epsilon > 0$. Then, there exists an absolute constant ϵ_0 such that, for all $r \in \mathbb{N}$,*

$$\mathbb{P} \left\{ \widetilde{\Omega}^r \right\} \geq 1 - \frac{\epsilon_0}{r^{p-2-\epsilon p}}.$$

The proof of Lemma C.6 appears in §F. Define

$$\mathfrak{X}^{r,x} = (\widehat{Z}^{r,x}, \widehat{W}^{r,x}, \widehat{X}^{r,x}, \widehat{Y}^{r,x}),$$

Re-define

$$E_r = \left\{ \mathfrak{X}^{r,x}(\omega, \cdot) : x \in \mathcal{X}^r \cap \{x : \|x\| \geq \delta r\}, \omega \in \Omega^r \right\}$$

and

$$\mathcal{E} = \{E_r, r \in \mathbb{N}\}.$$

Proof of Theorem C.4: To prove Theorem C.4 we must show that fixing $\epsilon, T > 0$, r large enough as well as $\omega \in \Omega^r$ and any $x \in \mathcal{X}^r \cap \{x : \|x\| \geq \delta r\}$, we have

$$\|\mathfrak{X}^{r,x}(\omega, \cdot) - \widetilde{\mathfrak{X}}(\cdot)\|_{s/\sqrt{r}, T} \leq \epsilon,$$

for some cluster point $\widetilde{\mathfrak{X}} = (\widetilde{Z}, \widetilde{W}, \widetilde{X}, \widetilde{Y})$ of \mathcal{E} with $\widetilde{\mathfrak{X}} \in E'$, and such that $(\widetilde{W}, \widetilde{Y})$ solves the SP with respect to $(\widetilde{W}(0), S, -R\gamma, R)$ with $\|\widetilde{W}(0)\| \leq \widetilde{\epsilon}$.

The existence of a cluster point that approximates $\widehat{X}^{r,x}$ in the above sense follows as in Proposition 4.1 of [5] and it remains to show that each cluster point satisfies the SP. To that end, note that, on $\widetilde{\Omega}^r$, we have

$$(1) \quad \|\widehat{W}^{r,x} - (\widehat{W}^{r,x}(0) + \widehat{X}^{r,x} + R\widehat{Y}^{r,x})\|_{s/\sqrt{r}, T} \leq \nu_\epsilon(r).$$

- (2) $\widehat{W}^{r,x}(t) \in \mathbb{R}_+^J$, and
 (3) (a) $\widehat{Y}^{r,x}(0) = 0$
 (b) $\widehat{Y}^{r,x}$ is nondecreasing, and
 (c) $\int_0^T \widehat{W}^{r,x}(t) d\widehat{Y}^{r,x}(t) = 0$.

Let $(\widetilde{Z}, \widetilde{W}, \widetilde{X}, \widetilde{Y})$ be a cluster point. Given properties (1)-(3b) above, it is obvious that properties (i) and (ii)(a)-(ii)(b) of the SP (see Definition 4) hold for $(\widetilde{W}, \widetilde{Y})$. The proof of the complementarity property (iii)(c) is identical to that in Proposition 6.2 of [5]. Finally, the fact that $\|\widetilde{W}(0)\| \leq \widetilde{\varepsilon}$ follows directly from the definition of $\widetilde{\Omega}_4^r$. ■

C.3 Completing the proof of Theorem 5.6

Proposition C.5 *Suppose that the conditions of Theorem 3.1 hold and fix $q < p$. Then, there exists an absolute constant $\kappa_q < \infty$ such that, for all $r \in \mathbb{N}$, $t > 0$, and $x \in \mathcal{X}^r$,*

$$\frac{1}{t} \int_0^t \mathbb{E}_x [\|Z^r(s)\|^q] ds \leq \kappa_q r^{\frac{5}{2}(q+1)} \left\{ \frac{1}{t} \|x\|^{q+1} + 1 \right\}.$$

In particular, for any $x \in \mathcal{X}^r$,

$$\mathbb{E}_{\pi^r} [\|Z^r(0)\|^q] = \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbb{E}_x [\|Z^r(s)\|^q] ds \leq \kappa_q r^{\frac{5}{2}(q+1)}.$$

With the uniform drift in Corollary C.5, to prove Proposition C.5 one repeats the proof of Theorem 5.5 in [16] almost verbatim. Care is needed, however, in identifying the dependence of the various constants on the heavy-traffic index r (in [16] the context is a given system and not a sequence of such). Towards that end, the following lemma, which is the adaptation of Proposition 5.3 in [16], is key. Here, we define

$$\tau_\delta^r(t_0) = \min \left\{ t \geq t_0 : \|\widehat{\Xi}^{r,x}(t)\| \leq \delta r \right\}.$$

Lemma C.7 *Fix $\delta > 0$. Then, for all $q < p$,*

$$\mathbb{E}_x \left[\int_0^{\tau_\delta^r(t_0)} (1 + \|\Xi^r(\sqrt{rt})\|^q) dt \right] \leq r^{\frac{3}{2}(q+1)}.$$

The proof of Lemma C.7 is relegated to §F.

Proof of Proposition C.5: We want to apply Theorem 5.5 in [16]. To that end, note that the constant $b^r > 0$ in Proposition 5.4 in [16] (where we add the superscript to make explicit the dependence on r) is set so that $b^r = \sup_{x \in C^r} V^r(x)$, where

$$V^r(x) = \mathbb{E}_x \left[\int_0^{\tau_\delta^r(t_0)} f(\Xi^r(\sqrt{rt})) dt \right],$$

and where, for the purposes of Theorem 5.5 the function $f(x) = 1 + \|x\|^q$ is used. The constant b^r is perturbed there to take care of the transition from discrete time to continuous time but for a crude bound it suffices to re-define $b^r = 2 \sup_{x \in C^r} V^r(x)$. It is then shown there that κ^r in the statement of Proposition 5.4 can be set to $\kappa^r = 3b^r$. By Lemma C.7 we have that $V^r(x) \leq c_0 r^{\frac{5}{2}(q+1)}$. Hence, we can set $\kappa^r = c_1 r^{\frac{5}{2}(q+1)}$.

The proof is completed by noting that the constant κ_p^r in [16] is given here by $\kappa_p^r = \kappa_q r^{\frac{5}{2}(q+1)}$ for an absolute constant κ_q . This analysis can be repeated for any $q < p$. Finally, the fact that

$$\mathbb{E}_{\pi^r} [\|Z^r(0)\|^q] = \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbb{E}_x [\|Z^r(s)\|^q] ds,$$

follows from Theorem 4.1 in [16]. ■

Appendix D. SSC via truncated fluid models In this section we prove Theorem 5.4. As before, we rely on [5] but will point out the places in which the use of the truncated fluid model entails a departure from [5]. Much of the section is dedicated to introducing the building blocks. The proof of Theorem 5.4 then appears at the end of the section.

We start by defining the hydrodynamically-scaled processes. Fix $L, \Theta = 3NL$ (with N as in (83)) and define

$$y^{r,m} = \|\epsilon^r(m\sqrt{r})\| \vee \sqrt{r}.$$

For all $r \in \mathbb{N}$, $m < \sqrt{r}T$, and $t \leq L$, define

$$\begin{aligned} Z^{r,m}(t) &= \frac{1}{y^{r,m}} (Z^r(m\sqrt{r} + y^{r,m}t) \wedge \Theta y^{r,m}), \\ W^{r,m}(t) &= \frac{1}{y^{r,m}} (W^r(m\sqrt{r} + y^{r,m}t) \wedge \Theta y^{r,m}), \\ \epsilon^{r,m}(t) &= \frac{1}{y^{r,m}} \epsilon^r(m\sqrt{r} + y^{r,m}t) \wedge \Theta y^{r,m}, \\ \mathcal{H}^{r,m}(t) &= \frac{1}{y^{r,m}} (\epsilon^r(m\sqrt{r} + y^{r,m}t) - \epsilon^r(m\sqrt{r})). \end{aligned}$$

We construct the processes $E^{r,m}, D^{r,m}, V^{r,m}, \Phi^{r,m}, T^{r,m}$ by applying the transformation $\mathfrak{D}^{y^{r,m}, \sqrt{r}}[\cdot]$, as in (82), to each of the processes E^r, D^r, Φ^r and T^r respectively. Finally,

$$(\varphi^k)^{r,m}(t) = \frac{1}{y^{r,m}} ((\varphi^k)^r(D_k^r(m\sqrt{r} + y^{r,m}t)) - (\varphi^k)^r(D_k^r(m\sqrt{r}))),$$

and we put

$$\mathfrak{X}^{r,m} = (A^{r,m}, D^{r,m}, T^{r,m}, W^{r,m}, Y^{r,m}, Z^{r,m}, \epsilon^{r,m}, \mathcal{H}^{r,m}).$$

By definition, there exists an absolute constant $\bar{\Theta}$ such that $\|\mathfrak{X}^{r,m}(0)\| \leq \bar{\Theta}$ for all $r, m \in \mathbb{N}$.

Thus far, a key departure from [5] is our use of a scalar that reflects only the initial value of $\epsilon^r = Z^r - \Delta W^r$ rather than that of Z^r (see equation (5.4) in [5]). The fact that we are able to proceed with the state-space collapse proofs with this “less informative” scaling owes to the special structure of the fluid models of queue-ratio network and to our requirement of linear attraction to the invariant manifold.

Fixing $\epsilon > 0$ and letting, as before, $\nu_\epsilon(r) = r^{-\epsilon}$, we define the following events on the underlying probability space:

$$\begin{aligned} \check{\Omega}_1^r &= \left\{ \max_{m < \sqrt{r}T} \|E^{r,m} - a^T \cdot\|_L \leq \nu_\epsilon(r) \right\}, \\ \check{\Omega}_2^r &= \left\{ \max_{m < \sqrt{r}T} \|D^{r,m} - M^{-1}T^{r,m}\|_L \leq \nu_\epsilon(r) \right\}, \\ \check{\Omega}_3^r &= \left\{ \max_{m < \sqrt{r}T} \left\| \sum_{k \in \mathcal{K}} (\varphi^k)^{r,m}(D^{r,m}) - \tilde{P}M^{-1}T^{r,m} \right\|_L \leq \nu_\epsilon(r) \right\}, \\ \check{\Omega}_4^r &= \left\{ \max_{m < \sqrt{r}T} \sup_{t_1, t_2 \leq L} \|\mathfrak{X}^{r,m}(t_2) - \mathfrak{X}^{r,m}(t_1)\| \leq \nu_\epsilon(r) + N|t_2 - t_1| \right\}. \end{aligned}$$

Let $\check{\Omega}^r = \bigcap_{i=1}^4 \check{\Omega}_i^r$, and define

$$u_k^{r,T,max} = \max\{|u_k^r(i)| : U_k^r(i-1) \leq rT, i = 1, 2, \dots\},$$

and

$$v_k^{r,T,max} = \max\{|v_k^r(i)| : V_k^r(i-1) \leq rT, i = 1, 2, \dots\}.$$

We next state probability bounds for the set $\check{\Omega}^r$. Lemma D.1 is the analogue of Lemma 5.1 in [5]. The sets \mathcal{B}_ϵ^r is as in (25).

Lemma D.1 *Fix $T > 0$ and $k \in \mathcal{K}$. Then, there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,*

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{P} \left\{ \frac{u_k^{r,T,max}}{\sqrt{r}} > \varepsilon \right\} \leq \frac{\varepsilon}{r^{p-3/2}}, \text{ and } \sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{P} \left\{ \frac{v_k^{r,T,max}}{\sqrt{r}} > \varepsilon \right\} \leq \frac{\varepsilon}{r^{p-3/2}}.$$

Proposition D.1 Fix $\epsilon, L, T > 0$. Then, there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,

$$\inf_{x \in \tilde{\mathcal{B}}_\epsilon^r} \mathbb{P}_x \{ \tilde{\Omega}^r \} \geq 1 - \varepsilon r^{-(p/2 - \epsilon p - 1)}.$$

The proofs of Lemma D.1 and Proposition D.1 are simplified versions of the proofs of Lemma C.3 and Proposition C.1 and are thus omitted. We next show that $\mathfrak{X}^{r,m}$ is approximated, in a proper sense, by a Lipschitz continuous function that satisfies the truncated-fluid-model equations. Let the family of functions E'_L be as defined in §C and let

$$E_r = \{ \mathfrak{X}^{r,m}(\omega, \cdot), m < \sqrt{r}T, \omega \in \tilde{\Omega}^r \}.$$

Note that each element $x \in E_r$ satisfies $\|x(0)\| \leq \bar{\Theta}$ as well as

$$\|x(t_2) - x(t_1)\| \leq N|t_2 - t_1| + \nu_\epsilon(r), \quad \text{for all } t_1, t_2 \in [0, L].$$

Put $\mathcal{E} = \{E_r, r \in \mathbb{N}\}$ and, as before, define a cluster point x of \mathcal{E} to be a point $x \in \mathcal{D}^d[0, L]$ such that for all $\epsilon > 0$ and $r^0 \in \mathbb{N}$, there exist $r \geq r^0$ and $y \in E_r$, with $\|x - y\|_L \leq \epsilon$.

The following is again the analogue of Proposition 6.1 in [5]. The proof is identical to the corresponding proof in [5] and is hence omitted. Below E'_L is as defined prior to Proposition C.2.

Proposition D.2 Fix $\delta, L, T > 0$. Then, for all r sufficiently large, $\omega \in \tilde{\Omega}^r$, and all $m < \sqrt{r}T$,

$$\|\mathfrak{X}^{r,m}(\omega, \cdot) - \tilde{\mathfrak{X}}(\cdot)\|_L \leq \delta,$$

for some cluster point $\tilde{\mathfrak{X}} = (\tilde{A}, \tilde{D}, \tilde{T}, \tilde{W}, \tilde{Y}, \tilde{Z}, \tilde{\epsilon}, \tilde{\mathcal{H}})$ of \mathcal{E} (that possibly depends on r and m) with $\tilde{\mathfrak{X}} \in E'_L$.

The following is the simple analogue of Proposition 8.1 in [5] adapted to our setting. The detailed proof is again omitted.

Lemma D.2 Any cluster point, $\tilde{\mathfrak{X}}$ of \mathcal{E} satisfies the truncated-fluid-model equations (16')-(22'). Moreover, $\tilde{\mathcal{H}}(t) = \tilde{\epsilon}(t) - \tilde{\epsilon}(0)$.

Recall that the linear attraction of the fluid model which is assumed in Theorem 5.4 implies that of the truncated fluid model. Thus, for each $\epsilon > 0$, there exists $t_0(\epsilon)$ such that $\|\tilde{\epsilon}(t_0)\| \leq \epsilon$. The following result is the analogue of Proposition 6.5 in [5].

Proposition D.3 Fix $\delta, T > 0$. Suppose that the conditions of Theorem 5.4 hold and that for each r , $\tilde{\Xi}^r(0) \in \mathcal{A}_\epsilon^r$. Then, there exist absolute constants L, ε (not depending on ϵ) so that for all sufficiently large r and all $\omega \in \tilde{\Omega}^r$,

$$\|\tilde{\epsilon}^r\|_{Ly^{r,0}/\sqrt{r}} \leq \varepsilon\epsilon,$$

and

$$\|\tilde{\epsilon}^r(t)\| \leq \delta(\|\tilde{\epsilon}^r\|_T \vee 1),$$

for all $t \in [Ly^{r,0}/\sqrt{r}, T]$ so that

$$\|\tilde{\epsilon}^r\|_{Ly^{r,0}/\sqrt{r}, T} \leq \delta.$$

The differences between Proposition D.3 and Proposition 6.5 in [5] follow from the fact that the scalar $y^{r,m}$ takes into account only the process $\tilde{\epsilon}^r$ whereas that of [5] has the workload. It is also important that $\tilde{\Xi}^r(0) \in \mathcal{A}_\epsilon^r$. With these minor differences taken into account, the proof of Proposition D.3 is similar to that of Proposition 6.5 in [5].

Proof of Theorem 5.4: The statements in (46) and (47) of the theorem follow directly from Proposition D.3 and it remains to establish (48). Note that

$$\|\tilde{\epsilon}^r(t)\| \leq \|\tilde{\epsilon}^r(0)\| + c_0 \frac{1}{\sqrt{r}} \left(\sum_{k \in \mathcal{K}} 1 + E_k^r(rt - \mathcal{R}_k^{a,r}(0) \wedge rt) + \sum_{k \in \mathcal{K}} 1 + S_k(rt - \mathcal{R}_k^{v,r}(0) \wedge rt) \right), \quad (90)$$

As in (72),

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\tilde{\epsilon}^r\|_T^q] \leq \epsilon + c_1 r^{\frac{q}{2}},$$

for any $q > 0$. Here we used the fact that $\|z - \Delta CMz\| \leq \epsilon$ for $x = (z, \varrho^a, \varrho^v) \in \mathcal{A}_\epsilon^r$. Using Proposition D.1 and Hölder's inequality we then obtain, for $q < p/2 - 1 - \epsilon p$ that,

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\tilde{\epsilon}^r\|_T^q \mathbb{1}\{\check{\Omega}^r\}^c] \leq c_3 \epsilon,$$

for an absolute constant c_3 that does not depend on ϵ and all sufficiently large r . By Proposition D.3 $\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\tilde{\epsilon}^r\|_T^q \mathbb{1}\{\check{\Omega}^r\}] \leq c_4 \epsilon$ for all sufficiently large $r \in \mathbb{N}$ and c_4 that does not depend on ϵ . In turn,

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\tilde{\epsilon}^r\|_T^q] \leq \sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\tilde{\epsilon}^r\|_T^q \mathbb{1}\{\check{\Omega}^r\}^c] + \sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\tilde{\epsilon}^r\|_T^q \mathbb{1}\{\check{\Omega}^r\}] \leq c_5 \epsilon,$$

where c_5 does not depend on ϵ . This concludes the proof of the Theorem. ■

Appendix E. State-space collapse in steady-state In this section we prove Theorem 5.3. In Theorem E.1 we establish a “downward drift” condition for the process $\epsilon^r = Z^r - \Delta W^r$ when Ξ^r is initialized in the set \mathcal{B}_ϵ^r . An application of the constrained-Lyapunov-function results, establishes a bound on the steady-state moments of ϵ^r . Corollary E.2 in this section proves Theorem 5.3 of the paper.

Below, for $x \in \mathbb{R}^K$, we let $\|x\|^h = \max_{1 \leq i \leq n} \langle h_i, x \rangle$ where h is the family of vectors in Definition 1.

Theorem E.1 *Suppose that the conditions of Theorem 3.1 hold and fix $\delta, \epsilon < 1/2$ and $\zeta > 1/2 - \epsilon$. Then, there exists an absolute constant t_0 such that, for all sufficiently large r ,*

$$\sup_{x \in \mathcal{B}_\epsilon^r: \|z - \Delta CMz\| > \delta r^\zeta} \mathbb{E}_x [\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h] \leq -\frac{\delta}{8} r^\zeta.$$

The proof of Theorem E.1 appears after the statement and proof of the following corollary.

Corollary E.2 *Suppose that the conditions of Theorem 3.1 hold. Then, for any $0 < \zeta < 1/2$ and $q < \frac{p}{3} + 1$ there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,*

$$\mathbb{E}_{\pi^r} [(\|\epsilon^r(0)\|^h)^{q-1}] \leq \varepsilon r^{q\zeta}. \quad (91)$$

In turn,

$$\mathbb{P}_{\pi^r} \{ \|\epsilon^r(0)\| > \epsilon \sqrt{r} \} \leq \frac{\varepsilon}{\epsilon^{q-1} r^{q(1/2-\zeta)}}.$$

The following will be used in the proof of Corollary E.2 and is proved in §F.

Lemma E.3 *Suppose that the conditions of Theorem 3.1 hold. Then, there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,*

$$\mathbb{P}_{\pi^r} \{ \widehat{\Xi}^r(0) \notin \mathcal{B}_\epsilon^r \} \leq \frac{\varepsilon}{r^{p(1/2-\epsilon)}}.$$

In passing, note that Lemma E.3 also proves (44).

Proof of Corollary E.2: Let $\widehat{\Xi}^{r,\zeta}(t) = \Xi^r(r^\zeta t)/r^\zeta$ with Ξ^r in (9). Namely, each of the components of Ξ^r is scaled in this manner. Let $\widehat{\mathcal{X}}^{r,\zeta}$ be the domain of the process $\widehat{\Xi}^{r,\zeta}$ and, for $x = (z, \varrho^a, \varrho^v) \in \widehat{\mathcal{X}}^{r,\zeta}$, define the function $\psi(\cdot) : \widehat{\mathcal{X}}^{r,\zeta} \rightarrow \mathbb{R}_+$ by

$$\psi(x) = 1 + \|z - \Delta CMz\|^h.$$

By Theorem E.1, $\psi(\cdot)$ is, for sufficiently large $r \in \mathbb{N}$, a constrained Lyapunov function of order $q = 1$ for the process $\widehat{\Xi}^{r,\zeta}$ with drift size parameter $-\delta/8$, drift-time parameter t_0 , exception parameter δ and inclusion set $\mathcal{B}_\epsilon^{r,\zeta}$ where

$$\mathcal{B}_\epsilon^{r,\zeta} = \left\{ x \in \widehat{\mathcal{X}}^{r,\zeta} : \|\varrho^a\| + \|\varrho^v\| \leq r^{1/2-\epsilon-\zeta}, k \in \mathcal{K} \right\};$$

see Definition 2.

We next apply Proposition 5.3 to show that $\psi(\cdot)$ is, in fact, a constrained Lyapunov function of order $q > 1$. Note that $\psi(\cdot) \geq 1$ so that, for all $q \geq 1$ and all x , $\psi^{q-2}(x) \leq \psi^{q-1}(x)$. Arguing as in the proof of Proposition B.1 one then has

$$|\psi^q(\widehat{\Xi}^{r,\zeta}(t_0)) - \psi^q(x)| \leq c_0 \psi^{q-1}(x) \left(|\psi(\widehat{\Xi}^{r,\zeta}(t_0)) - \psi(x)| + L_q^{\widehat{\Xi}^{r,\zeta}}(t_0, \mathcal{B}_\epsilon^{r,\zeta}) \right), \quad (92)$$

where

$$L_q^{\widehat{\Xi}^{r,\zeta}}(t_0, \mathcal{B}_\epsilon^{r,\zeta}) = \sup_{x \in \mathcal{B}_\epsilon^{r,\zeta}} \psi^{-(q-2)}(x) \mathbb{E}_x \left[(\psi(\widehat{\Xi}^{r,\zeta}(t_0)) - \psi(x))^2 (\psi(x) + |\psi(\widehat{\Xi}^{r,\zeta}(t_0)) - \psi(x)|)^{q-2} \right].$$

Since $\|\epsilon^r(r^\zeta t) - \epsilon^r(0)\| \leq c_1 \|E^r(r^\zeta t)\|$ we have by (72) that

$$c_0 := \limsup_{r \rightarrow \infty} L_q^{\widehat{\Xi}^{r,\zeta}}(t_0, \mathcal{B}_\epsilon^{r,\zeta}) < \infty$$

and, similarly that

$$c_2 := \limsup_{r \rightarrow \infty} \phi_q^{\widehat{\Xi}^{r,\zeta}}(t_0, \mathcal{A}) = \sup_{x \in \mathcal{B}_\epsilon^{r,\zeta}} \psi^{-(q-1)}(x) \mathbb{E}_x \left[(\psi^q(\widehat{\Xi}^{r,\zeta}(t_0)) - \psi^q(x))^+ \right] < \infty. \quad (93)$$

From Proposition 5.3 it then follows that $\psi(\cdot)$ is a constrained Lyapunov function of order q . Using (92) and similarly applying crude bounds, we have

$$\mathbb{E}_{\pi^r} \left[|\psi^q(\widehat{\Xi}^{r,\zeta}(t_0)) - \psi^q(\widehat{\Xi}^{r,\zeta}(0))| \mathbb{1}\{\widehat{\Xi}^{r,\zeta}(0) \notin \mathcal{B}_\epsilon^{r,\zeta}\} \right] \leq c_3 \mathbb{E}_{\pi^r} \left[\psi^{q-1}(\widehat{\Xi}^{r,\zeta}(0)) \mathbb{1}\{\widehat{\Xi}^{r,\zeta}(0) \notin \mathcal{B}_\epsilon^{r,\zeta}\} \right],$$

for all $r \in \mathbb{N}$. By Theorem 5.5, $\mathbb{E}_{\pi^r}[\psi^q(\widehat{\Xi}^{r,\zeta}(0))] < \infty$ (where π^r is the stationary distribution of $\widehat{\Xi}^r$). Thus, the conditions of Proposition 5.1 are satisfied and we conclude that

$$\mathbb{E}_{\pi^r} \left[\psi^{q-1}(\widehat{\Xi}^{r,\zeta}(0)) \right] \leq \left(1 + \frac{c_4}{\delta} \right) \mathbb{E}_{\pi^r} \left[\psi^{q-1}(\widehat{\Xi}^{r,\zeta}(0)) \mathbb{1}\{\widehat{\Xi}^r(0) \notin \mathcal{B}_\epsilon^r\} \right] + \frac{\kappa^{q-1} \phi_q^{\widehat{\Xi}^{r,\zeta}}(t_0, \mathcal{B}_\epsilon^{r,\zeta})}{\delta}, \quad (94)$$

for all $r \in \mathbb{N}$. Considering the first element on the right-hand side and applying Hölder's inequality, we have

$$\mathbb{E}_{\pi^r} \left[\psi^{q-1}(\widehat{\Xi}^{r,\zeta}(0)) \mathbb{1}\{\widehat{\Xi}^{r,\zeta}(0) \notin \mathcal{B}_\epsilon^{r,\zeta}\} \right] \leq \sqrt{\mathbb{E}_{\pi^r} \left[\psi^{2(q-1)}(\widehat{\Xi}^{r,\zeta}(0)) \right]} \sqrt{\mathbb{P}_{\pi^r} \left\{ \widehat{\Xi}^{r,\zeta}(0) \notin \mathcal{B}_\epsilon^{r,\zeta} \right\}}. \quad (95)$$

As $\psi(x) \leq c_5(1 + \|z\|)$ for all $x \in \widehat{\mathcal{X}}^{r,\zeta}$, it holds by Proposition C.5 that

$$\mathbb{E}_{\pi^r} \left[\psi^{2(q-1)}(\widehat{\Xi}^{r,\zeta}(0)) \right] \leq c_6 r^{\frac{3}{2}(2q-1)}, \quad (96)$$

so that, by Lemma E.3,

$$\limsup_{r \rightarrow \infty} \mathbb{E}_{\pi^r} \left[\psi^{q-1}(\widehat{\Xi}^{r,\zeta}(0)) \mathbb{1}\{\widehat{\Xi}^{r,\zeta}(0) \notin \mathcal{B}_\epsilon^{r,\zeta}\} \right] \leq c_7, \quad (97)$$

provided that $\frac{\delta}{2}(1/2 - \epsilon) \geq \frac{3}{4}(2q - 1)$. Using (93) and (97) in (94) concludes the proof of corollary. \blacksquare

The remainder of this section is dedicated to the proof of Theorem E.1. As we cannot directly take limits, we will apply the framework of Bramson [5] to a single time interval. To that end, let $y^r = r^\zeta$ and define

$$\mathfrak{X}^r(t) = \left(\frac{Z^r(y^r t)}{y^r} \wedge \Theta, \frac{W^r(y^r t)}{y^r} \wedge \Theta, \frac{\epsilon^r(y^r t)}{y^r} \wedge \Theta, \frac{T^r(y^r t)}{y^r}, \frac{\epsilon^r(y^r t) - \epsilon^r(0)}{y^r} \right).$$

As before, $\nu_\epsilon(r) = r^{-\epsilon}$ and we define the following subsets of Ω :

$$\begin{aligned}\tilde{\Omega}_1^r &= \left\{ \frac{1}{y^r} \|E^r(y^r \cdot) - a^r y^r \cdot\|_L \leq \nu_\epsilon(r) \right\}, \\ \tilde{\Omega}_2^r &= \left\{ \frac{1}{y^r} \|D^r(y^r \cdot) - M^{-1}T^r(y^r \cdot)\|_L \leq \nu_\epsilon(r) \right\}, \\ \tilde{\Omega}_3^r &= \left\{ \frac{1}{y^r} \left\| \sum_{k \in \mathcal{K}} (\varphi^k)^r (D^r(y^r \cdot)) - \tilde{P}M^{-1}T^r(y^r \cdot) \right\|_L \leq \nu_\epsilon(r) \right\}, \\ \tilde{\Omega}_4^r &= \left\{ \sup_{t_1, t_2 \leq L} \|\mathfrak{X}^r(y^r t_2) - \mathfrak{X}^r(y^r t_1)\| \leq \nu_\epsilon(r) + N|t_2 - t_1| \right\}.\end{aligned}$$

Let $\tilde{\Omega}^r = \bigcap_{i=1}^4 \tilde{\Omega}_i^r$. The following lemma is a simplified version of Proposition D.1 and we omit the proof.

Lemma E.4 *Fix $L, T, \epsilon > 0$. Then, there exists an absolute constant ε such that, for all $r \in \mathbb{N}$,*

$$\inf_{x \in \mathcal{B}_\epsilon^r} \mathbb{P}_x\{\tilde{\Omega}^r\} \geq 1 - \varepsilon r^{-(p/2 - \epsilon p - 1)}.$$

The notion of cluster points is a straightforward adaptation of its previous instances in this appendix and we do not repeat the relevant definitions. The proof of the following is a straightforward adaptation of Propositions 6.1 and 8.1 in [5].

Proposition E.1 *Fix $\delta, L, T > 0$. Suppose that $\widehat{\Xi}^r(0) \in \mathcal{B}_\epsilon^r$ for all $r \in \mathbb{N}$ and that the conditions of Theorem 5.3 hold. Then, for all sufficiently large $r \in \mathbb{N}$ and $\omega \in \tilde{\Omega}^r$,*

$$\|\mathfrak{X}^r(\omega, \cdot) - \tilde{\mathfrak{X}}(\cdot)\|_L \leq \frac{\delta}{4},$$

for some cluster point $\tilde{\mathfrak{X}} = (\tilde{Z}, \tilde{W}, \tilde{\epsilon}, \tilde{T}, \tilde{\mathcal{H}})$ of \mathcal{E} (that possibly depends on r) with $\tilde{\mathfrak{X}} \in E'$. Any such cluster point satisfies the truncated-fluid model equation (16') – (22'). Consequently, there exists t_0 such that if $\|\tilde{\epsilon}(0)\| \geq \delta$ then $\tilde{\mathcal{H}}(t_0) \leq -\delta/2$.

The following is an immediate corollary.

Corollary E.5 *Fix $\delta > 0$ and $0 < \zeta \leq 1/2$. Suppose that $\widehat{\Xi}^r(0) \in \mathcal{B}_\epsilon^r$ for all $r \in \mathbb{N}$ and that the conditions of Theorem 5.3 hold. Then, there exists an absolute constant t_0 such that, for all sufficiently large r and all $\omega \in \tilde{\Omega}^r$,*

$$\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h \leq -\frac{\delta}{2} r^\zeta \mathbb{1}\{\epsilon^r(0) \geq \delta r^\zeta\}.$$

We are now ready to prove Theorem E.1.

Proof of Theorem E.1: Fix $x \in \mathcal{B}_\epsilon^r$ with $\|\epsilon^r(0)\| \geq \delta r^\zeta$. Then, using Corollary E.5,

$$\begin{aligned}\mathbb{E}_x[\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h] &= \mathbb{E}_x[(\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h) \mathbb{1}\{\tilde{\Omega}^r\}] \\ &\quad + \mathbb{E}_x[(\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h) \mathbb{1}\{(\tilde{\Omega}^r)^c\}] \\ &\leq -\frac{\delta}{4} r^\zeta + \mathbb{E}_x[(\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h) \mathbb{1}\{(\tilde{\Omega}^r)^c\}].\end{aligned}$$

Since $\|\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h\| \leq c_0 \|E^r(r^\zeta t_0)\|$ the expectations above are finite. Using (72) together with Lemma E.4 and Hölder's inequality we conclude that

$$\sup_{x \in \mathcal{B}_\epsilon^r} \mathbb{E}_x[(\|\epsilon^r(r^\zeta t_0)\|^h - \|\epsilon^r(0)\|^h) \mathbb{1}\{(\tilde{\Omega}^r)^c\}] \leq \frac{\delta}{8}$$

for all sufficiently large $r \in \mathbb{N}$. Note that t_0 does not depend on the actual value of x beyond its being included in \mathcal{B}_ϵ^r . This concludes the proof of Theorem E.1. \blacksquare

Appendix F. Proofs of Auxiliary results In this section we prove the auxiliary results that were stated without proof in sections 2-5 of the paper and in §C-§E of this appendix. The proofs are divided into subsections according to the section of the paper in which the respective results appear.

F.1 Auxiliary results in sections 2-5

Proof of Lemma 2.1: Fix $L > 0$ and let $\bar{X} = (\bar{W}, \bar{Z}, \bar{T}, \bar{\epsilon}, \bar{\mathcal{H}})$ be a solution to the truncated fluid model equations (16')-(22') over $[0, L]$. We will prove that there exists a solution $\bar{X}^u = (\bar{W}^u, \bar{Z}^u, \bar{\epsilon}^u, \bar{T}^u)$ to the fluid model equations (16)-(22) over $[0, L]$ such that if \bar{X}^u is attracted to the invariant manifold at linear rate so is \bar{X} . Define $\bar{Z}^u(0) = \bar{Z}(0)$, (in turn $\bar{W}^u(0) = \bar{W}(0)$) and $\bar{\epsilon}^u(0) = \bar{\epsilon}(0)$. Define

$$\bar{T}^u(t) = \bar{T}(t) \text{ for all } t \in [0, L], \quad (98)$$

and construct $(\bar{Z}^u(t), \bar{W}^u(t), \bar{\epsilon}^u(t))$ using equations (16)-(18) and the definition (98). Equations (19) and (20) clearly hold by the definition of \bar{T}^u . To show that the constructed process \bar{X}^u satisfies (21) and (22) observe that if $\bar{Z}_k(0) \geq \Theta$ then, as $\Theta = 3NL$ (where N is the Lipschitz constant for \bar{X}), $\bar{T}(t)$ must satisfy that $\bar{Z}(0) + \alpha t - (I - \bar{P})M^{-1}\bar{T}(t) \geq 0$. In turn, also $\bar{Z}_k^u(t) > 0$ for all $t \in [0, L]$. If, $\bar{Z}_k(0) < \Theta$, then by construction $\bar{Z}_k^u(t) \leq \bar{Z}_k(t)$ for all $t \in [0, L]$. Thus, for $t \in [0, L]$, $\bar{W}_j^u(t) > 0$ implies $\bar{W}_j(t) > 0$ and, consequently, that $\int_0^L \bar{W}(s)d\bar{Y}(s) = 0$ so that \bar{X}^u satisfies (21). Similarly, if $\bar{\epsilon}_k(0) \geq \Theta$, then $\bar{\epsilon}_k^u(t) > 0$ for all $t \in [0, L]$ and, in any case, $\bar{\epsilon}_k^u(t) \leq \bar{\epsilon}_k(t)$ so that, for each $t \in [0, L]$, $\bar{\epsilon}_k(t) > 0$ implies $\bar{\epsilon}_k^u(t) > 0$ and we conclude that (22) holds for \bar{X}^u .

Since \bar{X}^u solves the fluid model equations (16)-(22) it is, by assumption, attracted to the invariant manifold at linear rate. By (18), (18a') and (98) it then holds that $\dot{\mathcal{H}}(t) = \dot{\bar{\epsilon}}^u(t)$ for all t such that $\dot{\bar{T}}(t) = \dot{\bar{T}}^u(t)$ exists, and, consequently, \bar{X} is attracted to the invariant manifold at linear rate. ■

Proof of equation (57): Using the fact that $\Phi(\cdot) > 1$ and arguing as in the proof of Proposition B.1 we have that

$$\mathbb{E}_x[|\Phi^q(\hat{\Xi}^r(t)) - \Phi^q(x)|] \leq c_0 \Phi^{q-1}(x) \left(\mathbb{E}_x[|\Phi(\hat{\Xi}^r(t)) - \Phi(x)|] + L_q^{\hat{\Xi}^r}(t) \right),$$

where $L_q^{\hat{\Xi}^r}(t)$ is as in (76). Recalling that $\Phi(x) = b + \Psi(CMz)$ for $x = (z, \varrho^a, \varrho^v)$ and using properties (P6) and (P7) of the function Ψ (see §5.2) we further have that $|\Phi(y) - \Phi(x)| \leq c_1 \|z_y - z_x\|$, for any $x = (z_x, \varrho_x^a, \varrho_x^v)$ and $y = (z_y, \varrho_y^a, \varrho_y^v)$ in $\hat{\mathcal{X}}^r$. In turn,

$$\mathbb{E}_x[|\Phi^q(\hat{\Xi}^r(t)) - \Phi^q(x)|] \leq c_2 \Phi^{q-1}(x) \mathbb{E}_x \left[\left(1 + \|\hat{Z}^r(t) - \hat{Z}^r(0)\| \right)^q \right].$$

By definition

$$\begin{aligned} \|\hat{Z}^r(t) - \hat{Z}^r(0)\| &\leq \frac{\|A^r(rt)\|}{\sqrt{r}} \\ &\leq c_3 \left(\frac{\sum_{k=1}^K (1 + E_k^r(rt - \mathcal{R}_k^{a,r}(0) \wedge rt))}{\sqrt{r}} + \frac{\sum_{k=1}^K (1 + S_k(rt - \mathcal{R}_k^{v,r}(0) \wedge rt))}{\sqrt{r}} \right). \end{aligned}$$

Observe that $E_k^r(rt - \mathcal{R}_k^{a,r}(0) \wedge rt)$ and $S_k(rt - \mathcal{R}_k^{v,r}(0) \wedge rt)$ are independent of $\hat{\Xi}^r(0)$. Equation (57) now follows from simple bounds for renewal processes as in (72). ■

Proof of Lemma 5.8: The proof requires only minor modifications to Lemmas 8.3 and 8.4 in [38]. The modifications reflect (and benefit from) our more restricted setting in which service times are generated upon commencement of service. Such a modification was carried out also in [1] and our arguments below are similar. Instead of repeating the proofs in [38] we underscore the points at which we depart from her proofs.

Two observations are important: (i) we will consider the two-parameters stopping time $(E^r(t), D^r(t))$ rather than $(E^r(t), A^r(t))$ as in [38], and (ii) we consider the two-parameter filtration

$$\mathcal{G}_{pq}^r = \sigma\{U_{\mathcal{K}^a}^r(\cdot \wedge (p + e_{\mathcal{K}^a})), V^r(\cdot \wedge (q + e)), \Phi^r(\cdot \wedge q), \hat{\Xi}^r(0)\}.$$

instead of the one defined in equation (92) of [38]. Above $e_{\mathcal{K}^a}$ denotes the vector of size $|\mathcal{K}^a|$ with all entries equal to 1. The following lemma is the analogue of Lemma 8.3 in [38]. The proof is identical except for obvious changes and it is omitted.

Lemma F.1 Fix $r \in \mathbb{N}$. The random time $\tau^r(t) = (E_{\mathcal{K}^a}^r(t), D^r(t))$ is a (multiparameter) stopping time relative to $\{\mathcal{G}_{pq}^r : (p, q) \in \mathbb{N}^{|\mathcal{K}^a|} \times \mathbb{N}^K\}$.

Next, we follow closely the argument in the proof of Lemma 8.4 in [38]. We define the processes \mathcal{M}^r and \mathcal{O}^r as in [38]; see equations (176) and (178) there. The process \mathcal{Q}^r (see equation (180) there) is defined differently in accordance with our different definition of the filtration \mathcal{G}_{pq}^r . Specifically, we let $\mathcal{Q}^{k,r} = \Phi^{k,r}(q_k) - \tilde{P}^k q_k$. Letting $\mathcal{T}^r(p, q) = (\mathcal{M}^r(p), \mathcal{O}^r(q), \mathcal{Q}^r(q))$ and repeating the exact arguments of [38] we may conclude that $((\mathcal{T}^r(\tau^r(t)), \mathcal{G}_{\tau^r(t)}^r, t \geq 0)$ is a martingale where $\tau^r(t) = (E_{\mathcal{K}^a}^r(t), D^r(t))$.

Define next

$$\check{\xi}^r(t) = \frac{1}{\sqrt{r}} C \left(\mathcal{O}^r(D^r(rt)) - M Q \mathcal{M}^r(E^r(rt)) + \sum_{k=1}^K \mathcal{Q}^{k,r}(D_k^r(rt)) \right),$$

and note that

$$\hat{\xi}^r(t) = \check{\xi}^r(t) + \hat{\zeta}^r(t),$$

where $\hat{\zeta}^r = C M \hat{\chi}^r$ and, for $k \in \mathcal{K}$,

$$\hat{\chi}_k^r(t) = \begin{cases} \frac{1}{\sqrt{r}} \left(\lambda_k^r (U_k^r(E^r(rt) + 1) - rt - u_k^r(1)) - (V_k^r(D_k^r(rt) + 1) - T^r(rt) - v_k^r(1)) \right), & \text{if } k \in \mathcal{K}^a, \\ -\frac{1}{\sqrt{r}} \left((V_k^r(D_k^r(rt) + 1) - T^r(rt) - v_k^r(1)) \right), & \text{otherwise.} \end{cases} \quad (99)$$

By the definition of the two-parameter stopping time, the process $\hat{\zeta}^r$ is adapted to the filtration $\mathcal{G}_{\tau^r(t)}^r$. So is the process Z^r (by the basic identities in (10) and (11)) and, in turn, the process W^r by the definition $W^r = C M Z^r$. The process Y^r is adapted to the filtration by the relation (13) and the fact that T^r is adapted follows from the construction in §2.2.

To conclude the proof it remains to establish the bound in item (ii) of the lemma. Note that (see e.g. equation (204) in [38]) that

$$\|\hat{\chi}_k^r\|_T \leq 2\lambda_k^r \frac{1}{\sqrt{r}} \sup_{0 \leq s \leq T} (u_k^r(E_k^r(rs) + 1)) + 2 \frac{1}{\sqrt{r}} \sup_{0 \leq s \leq T} (v_k^r(D_k^r(rs) + 1)),$$

where we recall that $u_k^r(i)$ is the i^{th} inter-arrival time for class k and $v_k^r(i)$ is the i^{th} service time. By definition $D_k^r(rt) \leq S_k(rt)$ and $\sup_{0 \leq s \leq T} |u_k^r(E_k^r(rs) + 1)| = \sup_{0 \leq s \leq T} |(u_k^r(E_k^r(rs) + 1))^q|$ (and similarly for v_k^r) so that

$$\mathbb{E}[\|\hat{\chi}_k^r\|_T^q]^{1/q} \leq \frac{1}{\sqrt{r}} \mathbb{E} \left[\sup_{0 \leq s \leq T} (u_k^r(E_k^r(rs) + 1))^q \right]^{1/q} + \frac{1}{\sqrt{r}} \mathbb{E} \left[\sup_{0 \leq s \leq T} (v_k^r(S_k(rs) + 1))^q \right]^{1/q}.$$

The bound for each of the expectations on the right-hand side now follows as in Lemma 8.4 of [17] applied to the stopped random walks with increments $(u_k(i))^q$ and $(v_k(i))^q$. This concludes the proof of the lemma. \blacksquare

Proof of Lemma 5.9: Throughout this proof the notation Δ represents the jump of a process rather than the lifting matrix. Recall that

$$\vartheta^r(t) = \sum_{s \leq t} \left(\Psi(\check{W}^r(s)) - \Psi(\check{W}^r(s-)) - \sum_{i=1}^J \partial_i \Psi(\check{W}^r(s-)) \Delta \check{W}_i^r(s) - \frac{1}{2} \sum_{i,l}^J \partial_i \partial_l \Psi(\check{W}^r(s-)) \Delta \check{W}_i^r(s) \Delta \check{W}_l^r(s) \right).$$

We treat separately the two parts of ϑ^r starting from the second line above. Recall that $\check{W}^r = \widehat{W}^r - R \widehat{\eta}^r + \hat{\zeta}^r$ where $\hat{\zeta}^r = C M \hat{\chi}^r$ and $\hat{\chi}^r$ is as defined in (99). Note that jumps of $\hat{\chi}_k^r$ have (up to a multiplicative

constant) the size of an interarrival time or a service time (scaled by \sqrt{r}) or the sum of such (in case of simultaneous jumps of E_k^r and D_k^r). Also, recall that $\widehat{W}^r = CM\widehat{Z}^r$ and $\widehat{\eta}^r(t) = CMQ\tilde{P}(\widehat{e}^r(t) - \widehat{e}^r(0))$ so that $\|\Delta(\widehat{W}^r(t) - R\widehat{\eta}^r(t))\| \leq c_0|\Delta\widehat{Z}^r(t)| \leq c_1/\sqrt{r}$.

Define, for each $j \in \mathcal{J}$,

$$v_j^r(t) = \sum_{k:s(k)=j} \left(\sum_{l=1}^{E_k^r(rt)+1} (u_k^r(l))^2 + \sum_{l=1}^{D_k^r(rt)+1} (v_k^r(l))^2 + (E_k^r(rt) + 1) + (D_k^r(rt) + 1) \right).$$

Then, for all $t \leq \tau_T^r$ and all $r \in \mathbb{N}$,

$$\begin{aligned} \left| \sum_{s \leq t} \sum_{i,l}^J \partial_i \partial_l \Psi(\check{W}^r(s-)) \Delta \check{W}_i^r(s) \Delta \check{W}_l^r(s) \right| &\leq \sup_{0 \leq s \leq \tau_T^r} \|D^2 \Psi(\check{W}^r(s))\| \frac{1}{r} \sum_{s \leq t} \sum_{i,l}^J \sqrt{r} |\Delta \check{W}_i^r(s)| \sqrt{r} |\Delta \check{W}_l^r(s)| \\ &\leq \epsilon \frac{1}{r} \sum_{s \leq t} \sum_{i,l}^J \sqrt{r} |\Delta \check{W}_i^r(s)| \sqrt{r} |\Delta \check{W}_l^r(s)| \leq \epsilon \frac{c_3}{r} \left(\sum_{i,l} \sqrt{v_i^r(T)} \sqrt{v_l^r(T)} \right). \end{aligned} \quad (100)$$

where we used the fact that $\|D^2 \Psi(\widehat{W}^r(s))\| \leq \epsilon$ for all $s \leq \tau_T^r$ by definition as well as Hölder's inequality. Similarly, by Taylor's theorem, we have for all $s \leq \tau_T^r$, that

$$\left| \Psi(\check{W}^r(s)) - \Psi(\check{W}^r(s-)) - \sum_{i=1}^J \partial_i \Psi(\check{W}^r(s-)) \Delta \check{W}_i^r(s) \right| \leq \|\Delta \check{W}^r(s)\|^2 \sup_{t \leq \tau_T^r} \|D^2 \Psi(\check{W}^r(s))\|,$$

so that

$$\left| \sum_{s \leq t} \Psi(\check{W}^r(s)) - \Psi(\check{W}^r(s-)) - \sum_{i=1}^J \partial_i \Psi(\check{W}^r(s-)) \Delta \check{W}_i^r(s) \right| \leq \epsilon \frac{c_4}{r} \sum_{i=1}^J v_i^r(T), \quad (101)$$

for all $t \leq \tau_T^r$, $r \in \mathbb{N}$. Combining (100) and (101) we conclude that for all $r \in \mathbb{N}$,

$$\begin{aligned} \|\vartheta^r\|_{\tau_T^r} &\leq \epsilon \frac{c_5}{r} \left(\sum_i v_i^r(T) + \sum_{i,l} \sqrt{v_i^r(T)} \sqrt{v_l^r(T)} \right) \\ &\leq \epsilon \frac{c_5}{r} \left(\sum_i \bar{v}_i^r(T) + \sum_{i,l} \sqrt{\bar{v}_i^r(T)} \sqrt{\bar{v}_l^r(T)} \right), \end{aligned}$$

with \bar{v}_i^r defined from v_i^r by replacing $D_k^r(rt)$ with $S_k(rt)$ and noting that, by definition, $D_k^r(rt) \leq S_k^r(rt)$. For $i \neq l$, $\bar{v}_i^r(T)$ and $\bar{v}_l^r(T)$ are independent random variables so that, using Jensen's inequality, we conclude that

$$\mathbb{E}[\|\vartheta^r\|_{\tau_T^r}] \leq \epsilon \frac{c_5}{r} \left(\sum_i \mathbb{E}[\bar{v}_i^r(T)] + \sum_{i,l} \sqrt{\mathbb{E}[\bar{v}_i^r(T)] \mathbb{E}[\bar{v}_l^r(T)]} \right). \quad (102)$$

Finally, since all service times and inter-arrival times have finite moments,

$$\mathbb{E} \left[\sum_{l=1}^{E_k^r(rt)+1} u_k^r(l) \right] = \mathbb{E}[u_k^r(2)] \mathbb{E}[E_k^r(rt) + 1] \text{ and } \mathbb{E} \left[\sum_{l=1}^{E_k^r(rt)+1} (u_k^r(l))^2 \right] = \mathbb{E}[(u_k^r(2))^2] \mathbb{E}[E_k^r(rt) + 1];$$

see e.g. the proof of Lemma 8.4 of [17]. Similarly,

$$\mathbb{E} \left[\sum_{l=1}^{S_k(rt)+1} v_k^r(l) \right] = \mathbb{E}[v_k^r(2)] \mathbb{E}[S_k(rt) + 1] \text{ and } \mathbb{E} \left[\sum_{l=1}^{S_k(rt)+1} (v_k^r(l))^2 \right] = \mathbb{E}[(v_k^r(2))^2] \mathbb{E}[S_k(rt) + 1].$$

By simple bounds for renewal processes (see equation (72)) we then have, for all $k \in \mathcal{K}$ and $r \in \mathbb{N}$, $\mathbb{E}[\bar{v}_k^r(T)] \leq c_6 + c_7 r T$. Plugging this back into (102) we have the result of the lemma. \blacksquare

Proof of Lemma 5.10: By property (P3) of $\Psi(\cdot)$ $c_0 := \sup_{w \in \mathbb{R}_+^J} \|D^2\Psi(w)\| < \infty$. Thus,

$$\sup_{0 \leq s \leq \tau_T^r} |D\Psi(\check{W}^r(s-)) - D\Psi(\widehat{W}^r(s-))| \leq c_0 \|\check{W}^r - \widehat{W}^r\|_{\tau_T^r} \leq c_0 (\|\zeta^r\|_{\tau_T^r} + \|R\hat{\eta}^r\|_{\tau_T^r}).$$

Next, recall that

$$\widehat{W}^r(t) = \widehat{X}^r(t) + R\widehat{Y}^r(t),$$

where \widehat{X}^r is as defined in (34). Also, for each $j \in \mathcal{J}$, \widehat{Y}_j^r is a non-negative process with nondecreasing sample paths that, together with \widehat{W}_j^r , satisfies

$$\int_0^t \mathbb{1}\{\widehat{W}_j^r(s) \in (0, \infty)\} d\widehat{Y}_j^r(s) = 0,$$

for all $t > 0$. By Theorem 5.1 in [39],

$$\sup_{0 \leq u \leq s \leq t} \|\widehat{Y}^r(s) - \widehat{Y}^r(u)\| \leq c_1 \sup_{0 \leq u \leq s \leq t} \|\widehat{X}^r(s) - \widehat{X}^r(u)\|,$$

for all $t \geq 0$. Since, \widehat{Y}^r is nondecreasing with $\widehat{Y}^r(0) = 0$ we have

$$\|\widehat{Y}^r\|_{\tau_T^r} = \|\widehat{Y}^r(\tau_T^r)\| \leq c_1 \sup_{0 \leq u \leq s \leq T} \|\widehat{X}^r(s) - \widehat{X}^r(u)\|.$$

Thus, for all $r \in \mathbb{N}$,

$$\sum_{i=1}^J \int_0^{t \wedge \tau_T^r} \left\| D\Psi(\check{W}^r(s-)) - D\Psi(\widehat{W}^r(s-)) \right\| \cdot |R^i| d\widehat{Y}_i^r(s) \leq c_2 (\|\zeta^r\|_{\tau_T^r} + \|R\hat{\eta}^r\|_{\tau_T^r}) \sup_{0 \leq u \leq s \leq t} \|\widehat{X}^r(s) - \widehat{X}^r(u)\|.$$

In turn,

$$\begin{aligned} & \sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\sum_{i=1}^J \int_0^{t \wedge \tau_T^r} \left\| D\Psi(\check{W}^r(s-)) - D\Psi(\widehat{W}^r(s-)) \right\| \cdot |R^i| d\widehat{Y}_i^r(s) \right] \\ & \leq c_3 \sqrt{\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\left(\sup_{0 \leq u \leq s \leq t} \|\widehat{X}^r(s) - \widehat{X}^r(u)\| \right)^2 \right]} \left(\sqrt{\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\|\zeta^r\|_{\tau_T^r}^2 \right]} + \sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\|R\hat{\eta}^r\|_{\tau_T^r}^2 \right] \right), \end{aligned} \quad (103)$$

for all $r \in \mathbb{N}$. Finally, recall that $\widehat{X}^r = \widehat{W}^r(0) + R(\widehat{\xi}^r + \widehat{\eta}^r)$ so that

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\left(\sup_{0 \leq u \leq s \leq t} \|\widehat{X}^r(s) - \widehat{X}^r(u)\| \right)^2 \right] \leq c_4 \left(\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\widehat{\xi}^r\|_T^2] + \sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x [\|\widehat{\eta}^r\|_T^2] \right) \leq c_5 \epsilon,$$

for all sufficiently large $r \in \mathbb{N}$ where c_5 does not depend on ϵ and the last inequality follows from Proposition A.1 and Theorem 5.4. Plugging this back into (103) and using item (ii) of Lemma 5.8 and Theorem 5.4 to bound η^r and ζ^r we get

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\sum_{i=1}^J \int_0^{t \wedge \tau_T^r} \left\| D\Psi(\check{W}^r(s-)) - D\Psi(\widehat{W}^r(s-)) \right\| \cdot |R^i| d\widehat{Y}_i^r(s) \right] \leq c_6 \epsilon,$$

with c_6 that does not depend on ϵ and all sufficiently large $r \in \mathbb{N}$. The second part of the lemma is proved similarly. This concludes the proof. \blacksquare

Proof of equations (66) and (67): Since $\|\check{W}^r - \widehat{W}^r\|_T \leq \|\zeta^r\|_T + \|R\hat{\eta}^r\|_T$ we can deduce from Lemma 5.11, Theorem 5.4 and item (ii) of Lemma 5.8 that, for all $r \in \mathbb{N}$,

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\left(|\Psi(\check{W}^r(t)) - \Psi(w)| \right)^q \right] \leq c_0. \quad (104)$$

Using Markov's inequality and property (P7) of the function $\Psi(\cdot)$ we have that

$$\sup_{x \in \mathcal{A}_\epsilon^r, \Psi(w) \geq 2\kappa} \mathbb{P}_x \{ \tau_T^r < T \} \leq \mathbb{P}_x \{ \|\check{W}^r - \check{W}^r(0)\|_T > \kappa \} \leq \frac{c_1}{\kappa},$$

for all $r \in \mathbb{N}$ and sufficiently large κ . By making κ larger if necessary this establishes (66). To show (67), note that, using properties (P6) and (P7) of the function $\Psi(\cdot)$, we have

$$\|\Psi(\check{W}^r(t)) - \Psi(\check{W}^r(t \wedge \tau_T^r))\| \leq c_2 \|\check{W}^r(t) - \check{W}^r(t \wedge \tau_T^r)\|.$$

Using (66) together with (104) and Hölder's inequality we have that

$$\sup_{x \in \mathcal{A}_\epsilon^r, \Psi(w) \geq 2\kappa} \mathbb{E}_x \left[\|\Psi(\check{W}^r(t)) - \Psi(\check{W}^r(t \wedge \tau_T^r))\| \right] \leq \frac{c_3}{\sqrt{\kappa}},$$

for all sufficiently large r and we can again re-choose a larger κ if necessary to obtain (67). \blacksquare

Proof of Lemma 5.11: By Theorem 5.1 in [39]

$$\sup_{0 \leq u \leq s \leq t} \|\widehat{W}^r(s) - \widehat{W}^r(u)\| \leq c_0 \sup_{0 \leq u \leq s \leq t} \|\widehat{X}^r(s) - \widehat{X}^r(u)\|,$$

and, in particular,

$$\|\widehat{W}^r(t) - \widehat{W}^r(0)\| \leq c_0 \sup_{0 \leq u \leq s \leq t} \|\widehat{X}^r(s) - \widehat{X}^r(u)\|.$$

Recall that $\widehat{X}^r(t) = \widehat{W}^r(0) + R(\widehat{\xi}^r(t) + \widehat{\eta}^r(t))$ where $\widehat{\eta}^r(t) = CMQ\widetilde{P}(\widehat{\mathcal{C}}^r(0) - \widehat{\mathcal{C}}^r(t))$. Applying Theorem 5.4 to bound $\widehat{\eta}^r$ and Proposition A.1 to bound $\widehat{\xi}^r$ we have, for all $r \in \mathbb{N}$, that

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\|\widehat{W}^r(t) - \widehat{W}^r(0)\|^q \right] \leq c_1. \quad (105)$$

Properties (P6) and (P7) of the function $\Psi(\cdot)$ (see §5.2) now imply that

$$\sup_{x \in \mathcal{A}_\epsilon^r} \mathbb{E}_x \left[\left(|\Psi(\widehat{W}^r(t)) - \Psi(\widehat{W}^r(0))| \right)^q \right] \leq c_2,$$

for all $r \in \mathbb{N}$ as required. ■

F.2 Auxiliary results in §C

Proof of Corollary C.2: The proof mimics the proof of Proposition 8.1 in [38]. Recall that

$$\widehat{W}^{r,x}(t) = \widehat{X}^{r,x}(t) + R\widehat{Y}^{r,x}(t),$$

where $\widehat{X}^{r,x}(t) = \widehat{W}^{r,x}(0) + R(\widehat{\xi}^{r,x}(t) + \widehat{\eta}^{r,x}(t))$, with $\widehat{\xi}^{r,x}$ and $\widehat{\eta}^{r,x}$ as defined in (89). Also, for each $j \in \mathcal{J}$, $\widehat{Y}_j^{r,x}$ is a non-negative process with nondecreasing sample paths that, together with $\widehat{W}_j^{r,x}$, satisfies that

$$\int_0^t \mathbb{1}\{\widehat{W}_j^{r,x}(s) \in (0, \infty)\} d\widehat{Y}_j^{r,x}(s) = 0,$$

for each $t > 0$. By Theorem 5.1 in [39],

$$\sup_{\frac{s}{\sqrt{r}} \leq u \leq t \leq T} \|\widehat{W}^{r,x}(t) - \widehat{W}^{r,x}(u)\| \leq c_0 \sup_{\frac{s}{\sqrt{r}} \leq u \leq t \leq T} \|\widehat{X}^{r,x}(t) - \widehat{X}^{r,x}(u)\|.$$

In turn, for all $r \in \mathbb{N}$,

$$\begin{aligned} \|\widehat{W}^{r,x}\|_{\frac{s}{\sqrt{r}}, T} &\leq \left\| \widehat{W}^{r,x} \left(\frac{s}{\sqrt{r}} \right) \right\| + c_1 \|\widehat{X}^{r,x}\|_T \\ &\leq \left\| \widehat{W}^{r,x} \left(\frac{s}{\sqrt{r}} \right) \right\| + c_1 \|\widehat{\xi}^{r,x}\|_{\frac{s}{\sqrt{r}}, T} + c_1 \sup_{\frac{s}{\sqrt{r}} \leq u \leq T} \left\| \widehat{\eta}^{r,x}(u) - \widehat{\eta}^{r,x} \left(\frac{s}{\sqrt{r}} \right) \right\| \\ &\leq \left\| \widehat{W}^{r,x} \left(\frac{s}{\sqrt{r}} \right) \right\| + c_1 \|\widehat{\xi}^{r,x}\|_T + c_2 (\|\widehat{W}^{r,x}\|_T \vee 1) \frac{\|\widehat{\mathcal{C}}^{r,x}\|_{\frac{s}{\sqrt{r}}, T}}{\|\widehat{W}^{r,x}\|_T \vee 1} \end{aligned} \quad (106)$$

where the last inequality follows from the definition $\widehat{\eta}^{r,x}(t) = CMQ\widetilde{P}(\widehat{\mathcal{C}}^{r,x}(t) - \widehat{\mathcal{C}}^{r,x}(0))$. Since $\|\widehat{Z}^{r,x}(t)\| \leq \|\widehat{Z}^{r,x}(0)\| + \sum_k E_k^r(\|x\|\sqrt{rt})/\|x\|$, we have $\|\widehat{W}^{r,x}\|_{\frac{s}{\sqrt{r}}} \leq \|\widehat{W}^{r,x}(0)\| + c_3 \|E^r(\|x\|s)\|/\|x\|$. Also, $\widehat{W}^{r,x}(0) \leq 1$ by definition. Thus,

$$\|\widehat{W}^{r,x}\|_{\frac{s}{\sqrt{r}}} \leq 1 + c_3 \left(1 + \frac{\|E^r(\|x\|s)\|}{\|x\|} \right) \text{ and } \|\widehat{W}^{r,x}(0)\| \leq 1.$$

On the set Ω^r (as defined prior to Proposition C.1), we have that $\|E^r(\|x\|s)\|/\|x\| \leq c_5$. By Proposition C.4, we have, on Ω^r , that $\|\widehat{\mathcal{C}}^{r,x}\|_{\frac{s}{\sqrt{r}}, T}/(\|\widehat{W}^{r,x}\|_T \vee 1) \leq \epsilon$. Plugging these into (106) we have that, on Ω^r ,

$$\|\widehat{W}^{r,x}\|_T \leq c_6 + c_7 \|\widehat{\xi}^{r,x}\|_T + c_7 \epsilon (\|\widehat{W}^{r,x}\|_T \vee 1).$$

In turn, $\|\widehat{W}^{r,x}\|_T \leq 2(c_6 + c_7 \|\widehat{\xi}^{r,x}\|_T)$ on Ω^r . Using the probability bound on Ω^r in Proposition C.1 we get

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{\|\widehat{\mathcal{C}}^{r,x}\|_{\frac{s}{\sqrt{r}}, T}}{c_6 + c_7 \|\widehat{\xi}^{r,x}\|_T} > \epsilon \right\} \leq \frac{c_8}{r^{p-2-\epsilon p}}.$$

Finally, from Lemma A.1 it is easily verified that $\mathbb{P}\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|\hat{\xi}^{r,x}\|_T \geq 1\} \leq c_9 r^{-q}$ for all $q < p-1$. We conclude that

$$\begin{aligned} \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|\hat{e}^{r,x}\|_{\frac{s}{\sqrt{r}}, T} > \epsilon(c_6 + c_7)\right\} &\leq \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{\|\hat{e}^{r,x}\|_{\frac{s}{\sqrt{r}}, T}}{c_6 + c_7 \|\hat{\xi}^{r,x}\|_T} > \epsilon\right\} \\ &\quad + \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|\hat{\xi}^{r,x}\|_T \geq 1\right\} \\ &\leq \frac{c_{10}}{r^{p-2-\epsilon p}}, \end{aligned}$$

for all $r \in \mathbb{N}$. ■

Proof of Lemma C.3: We fix $k \in \mathcal{K}$ and omit it from the notation. We write the proof only for $u_k^{r,x,T,max}$. The proof for $v_k^{r,x,T,max}$ is similar.

For all sufficiently large r ,

$$\begin{aligned} \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{u^{r,x,T,max}}{\|x\|} > \epsilon\right\} &\leq \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{E^r(\|x\|\sqrt{r}T)}{\|x\|\sqrt{r}} > 2\alpha^r T\right\} \\ &\quad + \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{\max_{2 \leq i \leq 3\alpha\|x\|\sqrt{r}T} u^r(i)}{\|x\|} > \epsilon\right\} \end{aligned} \quad (107)$$

where we recall that α^r is the rate of the renewal process E^r and $\alpha = \lim_{r \rightarrow \infty} \alpha^r$.

We next bound each of the components on the right-hand side of (107). Note that

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{E^r(\|x\|\sqrt{r}T)}{\|x\|\sqrt{r}} > 2\alpha^r T\right\} = \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{E^r(\|x\|\sqrt{r}T) - \alpha^r \|x\|\sqrt{r}T}{\|x\|\sqrt{r}} > \alpha^r T\right\},$$

Using equation (69) (with x there replaced by $\alpha^r \|x\|\sqrt{r}T$ and rT replaced by $x\sqrt{r}T$), we then have that

$$\sup_{x \in \mathcal{X}^r: \|x\| > \delta r} \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{E^r(\|x\|\sqrt{r}T)}{\|x\|\sqrt{r}} > 2\alpha^r T\right\} \leq \frac{c_0}{r^{p-1}},$$

where c_0 may depend on δ .

For the second element on the right-hand side of (107), note that

$$\begin{aligned} \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{\max_{2 \leq i \leq 3\alpha\|x\|\sqrt{r}T} u^r(i)}{\|x\|} > \epsilon\right\} &\leq \sum_{k=1}^{\infty} \mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \delta kr \|x\| \leq \delta(k+1)r} \frac{\max_{2 \leq i \leq 3\alpha\|x\|\sqrt{r}T} u^r(i)}{\|x\|} > \epsilon\right\} \\ &\leq \sum_{k=1}^{\infty} \mathbb{P}\left\{\frac{\max_{2 \leq i \leq 3\alpha\delta(k+1)r\sqrt{r}T} u^r(i)}{\delta kr} > \epsilon\right\} \end{aligned}$$

Then, by Markov inequality

$$\mathbb{P}\left\{\frac{\max_{1 \leq i \leq 3\alpha\delta(k+1)r\sqrt{r}T} u^r(i)}{\delta kr} > \epsilon\right\} \leq 1 - \left(1 - \frac{\mathbb{E}[(u^r(2))^p]}{(\epsilon\delta kr)^p}\right)^{[\delta(k+1)r^{3/2}T]} \leq c_1 \frac{1}{(kr)^{p-3/2}}, \quad (108)$$

for all $r \in \mathbb{N}$ where the last inequality follows from simple manipulations. Summing over $k = 1, 2, \dots$, we have, for all $r \in \mathbb{N}$,

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{\max_{1 \leq i \leq 3\alpha\|x\|\sqrt{r}T} u^r(i)}{\|x\|} > \epsilon\right\} \leq \frac{c_2}{r^{p-3/2}}$$

To prove the moment bounds note that, by definition,

$$\frac{u^{r,x,T,max}}{\|x\|} \leq \sqrt{r}T + \frac{u^r(E^r(\|x\|\sqrt{r}T) + 1)}{\|x\|}.$$

Since $\sup_{r \rightarrow \infty} \mathbb{E}[(u^r(1))^q] < \infty$, fixing sufficiently large r , we have that $\mathbb{E}[(u^r(E^r(\|x\|\sqrt{r}T) + 1))^q] \leq c_0(\|x\|\sqrt{r}T)$; see e.g. [21, Theorem 2.6.3]. Consequently

$$\sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \mathbb{E} \left[\left(\frac{u^{r,x,T,max}}{\|x\|} \right)^q \right] \leq c_3 \sqrt{r}T.$$

The moment bound now follows from the probability bounds through an application of Hölder's inequality. \blacksquare

Proof of Proposition C.1: We consider first the set Ω_1^r . Note that, since $\mathcal{R}_k^{a,r}(0) \leq \|x\|$ (where $x = \Xi^r(0)$) we have for $m \geq 1$ that $\mathcal{R}_k^{a,r}(m\|x\|) \leq u_k^{r,x,T,max}$. In turn,

$$\begin{aligned} & \mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|E^{r,m,x} - \alpha^r \cdot\|_L > \nu_\epsilon(r) \right\} \leq \mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \frac{u^{r,x,T,max}}{\|x\|} > \zeta \right\} \\ & \quad + \mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|1 + \tilde{E}^r(y^{r,m,x}t) - \alpha^r y^{r,m,x}t\|_{L-\zeta} > \nu_\epsilon(r) y^{r,m,x} - 1 \right\} \\ & \leq \frac{c_0}{r^{p-3/2-\epsilon p}}. \end{aligned}$$

Here, \tilde{E}^r is the undelayed version of E^r , i.e., one in which the first inter-arrival time has the same distribution as all subsequent inter-arrival times. Lemma C.3 bounds the first element on the right hand side while the bound for the second element follows as in the beginning of the proof of Lemma C.3. Thus,

$$\mathbb{P} \left\{ \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|E^{r,m,x} - \alpha^r \cdot\|_L > \nu_\epsilon(r) \right\} \leq \frac{c_0}{r^{p-3/2-\epsilon p}},$$

for all $r \in \mathbb{N}$. Multiplying this bound by the number of intervals $\lfloor \sqrt{r}T \rfloor$ we get

$$\mathbb{P} \{ (\Omega_1^r)^c \} = \mathbb{P} \left\{ \max_{m < \sqrt{r}T} \sup_{x \in \mathcal{X}^r: \|x\| \geq \delta r} \|E^{r,m,x} - \alpha^r \cdot\|_L > \nu_\epsilon(r) \right\} \leq \frac{c_1}{r^{p-2-\epsilon p}}.$$

The bounds for Ω_i^r for $i = 2, 3$ follow similarly. Given the bounds in Lemma C.3, the proof for Ω_4^r follows closely the proof of Proposition 5.2 in [5] and the details are omitted. \blacksquare

Proof of Lemma C.6: The bound for $\tilde{\Omega}_1^r$ follows from the definition $\hat{X}^{r,x}$, from Theorem C.2 and applying Lemma A.1 to bound the centered and scaled renewal processes.

The bounds for $\tilde{\Omega}_2^r$ and $\tilde{\Omega}_3^r$ follow by invoking Assumption 1 and applying the Oscillation inequalities as in the proof of Corollary C.2. To that end, recall that for each $r \in \mathbb{N}$ and $x \in \mathcal{X}^r$,

$$\widehat{W}^{r,x}(t) = \hat{X}^{r,x}(t) + R\hat{Y}^{r,x}(t),$$

where $\hat{X}^{r,x}(t) = \widehat{W}^{r,x}(0) + R(\hat{\xi}^{r,x}(t) + \hat{\eta}^{r,x}(t))$, with $\hat{\xi}^{r,x}$ and $\hat{\eta}^{r,x}$ as defined in (89). Also, for each $j \in \mathcal{J}$, $\hat{Y}_j^{r,x}$ is a non-negative process with nondecreasing sample paths that, together with $\widehat{W}_j^{r,x}$, satisfies, for all $t \geq 0$, that

$$\int_0^t \mathbb{1}\{\widehat{W}_j^{r,x}(s) \in (0, \infty)\} d\hat{Y}_j^{r,x}(s) = 0,$$

for each $t > 0$. By Theorem 5.1 in [39],

$$\sup_{0 \leq u \leq s \leq t} \|\widehat{W}^{r,x}(s) - \widehat{W}^{r,x}(u)\| \leq c_1 \sup_{0 \leq u \leq s \leq t} \|\hat{X}^{r,x}(s) - \hat{X}^{r,x}(u)\|,$$

and

$$\sup_{0 \leq u \leq s \leq t} \|\hat{Y}^{r,x}(s) - \hat{Y}^{r,x}(u)\| \leq c_1 \sup_{0 \leq u \leq s \leq t} \|\hat{X}^{r,x}(s) - \hat{X}^{r,x}(u)\|.$$

In turn, the bounds for $\tilde{\Omega}_2^r$ and $\tilde{\Omega}_3^r$ follow from that for $\tilde{\Omega}_1^r$. Finally, the bound on $\tilde{\Omega}_4^r$ follows from crude bound on the arrivals as in the proof of Corollary C.2. \blacksquare

Proof of Lemma C.7: We add the superscript r to make explicit the dependence on r , but otherwise try to follow the notation in [16]. First, we have $L^r = \delta r$ so that the compact set we consider is $C^r = \{x \in \mathcal{X}^r : \|x\| \leq \delta r\}$. Also, we let $t^r(x) = t_0 \sqrt{r} \max(L^r, \|x\|)$ where t_0 is as in Corollary C.5. Note that $t^r(x) \geq t_0 \delta r^{3/2}$.

Focusing on the proof of Proposition 5.3 in [16] note that the constant $b > 0$ on the right-hand side of (5.5) in [16] is obtained by bounding in a crude manner the behavior when $x \notin C^r$. In that case $\mathbb{E}_x[\|\Xi^r(t^r(x))\|^{p+1}] \leq \|x\|^{p+1} + \tilde{c} \mathbb{E}_x[\|E^r(t^r(x))\|]$ for some constant \tilde{c} . Applying simple renewal bounds (as in (72)) in the definition of $t^r(x)$ there we may replace b with $b^r = \tilde{c} r^{\frac{3}{2}(p+1)}$ for a re-defined absolute constant \tilde{c} (note that \tilde{c} is indeed an absolute constant – one that does not change with r). Following into the next paragraph in [16], we set $\sigma_1^r = t^r(x)$ and for all $k = 2, \dots$, and $\sigma_{k+1}^r = \sigma_k^r + \theta_{\sigma_k^r} \sigma_1^r$. Note that σ_k^r depends on the initial condition x .

The next step is to identify the appropriate value (as a function of r) for the constant c_0^r on the right hand side of (5.7) in [16]. This constant is identified by examining closely the constants $c_1 - c_6$ there as follows:

- (i) The constant c_1 does not depend on r and we may hence set $c_1^r \equiv \tilde{c}_1$.
- (ii) The constant c_2^r is bounded by $\check{c}_2(\sigma_1^r)^2$ for an absolute constant \check{c}_2 and, consequently, we may set $c_2^r = \tilde{c}_2 r^{3/2}$ for an absolute constant \tilde{c}_2 .
- (iii) Similarly to c_2^r , we can set $c_3^r = \tilde{c}_3 r^{3/2}$ for an absolute constant \tilde{c}_3 .
- (iv) The constant c_4^r does not change with r so that we may set $c_4^r \equiv \tilde{c}_4$ for an absolute constant \tilde{c}_4 .
- (v) The same follows for c_5^r and we may set $c_5^r \equiv \tilde{c}_5$ for an absolute constant \tilde{c}_5 .
- (vi) The constant c_6^r is bounded by $\check{c}_6(\sigma_1^r)^{p+1}$ for some absolute constant \check{c}_6 and in turn, we can set $c_6^r = \tilde{c}_6 r^{\frac{3}{2}(p+1)}$ for an absolute constant \tilde{c}_6 .

Plugging the above back into equations (5.7) and (5.8) in [16] we obtain that c_0^r there can be replaced with $c_0^r = \tilde{c}_0 r^{\frac{3}{2}(p+1)}$ for an absolute constant \tilde{c}_0 . Plugging this constant into (5.6) there, we obtain that the constant c_7 in the last display of the proof of Proposition 5.3 can be replaced by $c_7^r = \tilde{c}_7 r^{\frac{3}{2}(p+1)} b^r (\|x\|^{p+1} + 1)$, for an absolute constant \tilde{c}_7 . This also replaces the constant in the statement of that proposition. Namely, we set $c_{p+1}^r = \tilde{c}_7 r^{\frac{3}{2}(p+1)} b^r = \tilde{c}_8 r^{3(p+1)}$ for an absolute constant \tilde{c}_8 .

The statement of the lemma now follows from Proposition 5.3 in [16] with the above replacements of constants. ■

F.3 Auxiliary results in §E

Proof of Lemma E.3: The proof is an adaptation of the proof of Lemma 3.2 in [26] but we repeat the relevant portion for completeness.

$$\begin{aligned}
\mathbb{P}_{\pi^r} \{\widehat{\mathcal{R}}_k^{v,r}(0) > r^{-\epsilon}\} &= \mathbb{P}_{\pi^r} \{\mathcal{R}_k^{v,r}(0) > r^{1/2-\epsilon}\} \\
&= \sum_{z \in \mathbb{Z}_+^K} \int_{e_k^v \in (0, \infty)} \mathbb{P}_{\pi^r} \{\mathcal{R}_l^{v,r}(0) > r^{1/2-\epsilon} | Z_l = z_l\} \cdot \mathbb{P}_{\pi^r} \{Z_l = z_l\} \\
&\leq \sum_{z \in \mathbb{Z}_+^K} \mathbb{P}_{\pi^r} \{Z_l = z_l\} \sup_{z \in \mathbb{R}_+} \frac{1 - F_k^s(r^{1/2-\epsilon} + z)}{1 - F_k^s(z)} \\
&\leq \sup_{z \in \mathbb{R}_+} \frac{\mathbb{E}[(v_k(2) - z)^p | v_k(2) > z]}{r^{1/2-\epsilon}} \leq \frac{c_0}{(r^{1/2-\epsilon})^p},
\end{aligned}$$

where, in the last line, we used Markov's inequality and condition (4).

The proof for the residual inter-arrival can be done using the equilibrium distribution of the renewal process as in [26] to get that

$$\begin{aligned}
\mathbb{P}_{\pi^r} \{\mathcal{R}_k^{a,r}(0) > r^{1/2-\epsilon}\} &= \alpha_k^r \int_{r^{1/2-\epsilon}}^{\infty} 1 - F_k^{a,r}(x) dx \leq \mathbb{E}[u_k^r(2) \mathbb{1}\{u_k^r(2) > r^{1/2-\epsilon}\}] \\
&\leq \sqrt{\mathbb{E}[(u_k^r(2))^2] (1 - F_k^{a,r}(r^{1/2-\epsilon}))} \leq \sqrt{\mathbb{E}[(u_k^r(2))^2]} \frac{\mathbb{E}[(u_k^r(2))^p]}{(r^{1/2-\epsilon})^p}
\end{aligned}$$

where we have used first Hölder's inequality and then Markov's inequality. ■

Acknowledgments. The author wishes to thank Assaf Zeevi for his significant contribution to this manuscript. Proposition B.1 and its proof are due to David Gamarnik and Assaf Zeevi.

References

- [1] B. Ata and W. Lin, *Heavy traffic analysis of maximum pressure policies for stochastic processing networks with multiple bottlenecks*, Queueing Systems **59** (2008), no. 3, 191–235.
- [2] A. Bernard and A. El Kharroubi, *Régulation de processus dans le premier orthant de \mathbb{R}^n* , Stochastics and Stochastics Rep. **34** (1991), 149–167.
- [3] M. Bramson, *Two badly behaved queueing networks*, Stochastic Networks (F. Kelly and R. Williams, eds.), vol. 71, Proceedings of the IMA, 1995, pp. 105–116.
- [4] _____, *Convergence to equilibria for fluid models of FIFO queueing networks*, Queueing Systems **22** (1996), 5–45.
- [5] _____, *State space collapse with applications to heavy-traffic limits for multiclass queueing networks*, Queueing Systems **30** (1998), 89–148.
- [6] A. Budhiraja and C. Lee, *Stationary distribution convergence for generalized Jackson networks in heavy traffic*, Mathematics of Operations Research **34** (2009), no. 1, 45–56.
- [7] A. Budhiraja and C. Lee, *Long time asymptotics for constrained diffusions in polyhedral domains*, Stoch. Proc. and their App. **117** (2007), 1014–1036.
- [8] H. Chen, *A sufficient condition for the positive recurrence of a semimartingale reflecting Brownian motion in an orthant*, Ann. Appl. Prob. **6** (1996), no. 3, 758–765.
- [9] H. Chen and X. Shen, *Computing the stationary distribution of an SRBM in an orthant with applications to queueing networks*, Queueing Systems **45** (2003), 27–45.
- [10] H. Chen and H.Q. Ye, *Existence condition for the diffusion approximation of multiclass priority queueing networks*, Queueing Systems **38** (2001), 435–470.
- [11] H. Chen and H. Zhang, *Diffusion approximations for re-entrant lines with a first-buffer-first-served priority discipline*, Queueing Systems **23** (1996), 177–195.
- [12] _____, *A sufficient condition and a necessary condition for the diffusion approximations of multi-class queueing networks under priority service disciplines*, Queueing Systems **34** (2000), 237–268.
- [13] M. Csörgö and L. Horváth, *Weighted approximations in probability and statistics*, John Wiley & Sons; Chichester, 1996.
- [14] J. G. Dai, *On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Prob. **5** (1995), 49–77.
- [15] J. G. Dai and J.M. Harrison, *Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis*, Ann. Appl. Prob. **2** (1992), 65–86.
- [16] J. G. Dai and S. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Aut. Control **40** (1995), 1889–1904.
- [17] JG Dai and W. Dai, *A heavy traffic limit theorem for a class of open queueing networks with finite buffers*, Queueing Systems **32** (1999), no. 1, 5–40.
- [18] P. Dupuis and R. J. Williams, *Lyapunov function for semimartingale reflecting Brownian motions*, Annals of Probability **22** (1994), 680–702.
- [19] D. Gamarnik and A. Stolyar, *Multiclass multiserver queueing system in the halfin-whitt heavy traffic regime. asymptotics of the stationary distribution*, Arxiv preprint arXiv:1105.0635 (2011).
- [20] D. Gamarnik and A. Zeevi, *Validity of heavy traffic steady-state approximations in generalized Jackson networks*, Ann. Appl. Prob. **16** (2006), 56–90.
- [21] A. Gut, *Stopped random walks: limit theorems and applications*, Springer Verlag, 2009.

- [22] J. M. Harrison, *The heavy traffic approximation for single server queues in series*, Journal of Appl. Prob. **10** (1973), 613–629.
- [23] J. M. Harrison and V. Nguyen, *Brownian models of multiclass queueing networks: Current status and open problems*, Queueing Systems **13** (1993), 5–40.
- [24] D. L. Iglehart and W. Whitt, *Multiple channel queues in heavy traffic. i*, Advances in Applied Probability **2** (1970), 150–177.
- [25] WN Kang, FP Kelly, NH Lee, and RJ Williams, *State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy*, The Annals of Applied Probability **19** (2009), no. 5, 1719–1780.
- [26] T. Katsuda, *State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic*, Queueing Systems **65** (2010).
- [27] J.F.C Kingman, *The single server queue in heavy traffic*, Proc. Camb. Phil. Soc. **57** (1961).
- [28] A. Mandelbaum and S. Stolyar, *Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$ -rule*, Oper. Res. **52** (2004), 836–855.
- [29] A. Mandelbaum and A. Van der Heyden, *Complementarity and reflection*, Unpublished work, 1987.
- [30] W. P. Peterson, *Diffusion approximations for networks of queues with multiple customer types*, Math. Oper. Res. **9** (1991), 90–118.
- [31] M. I. Reiman, *Open queueing networks in heavy traffic*, Math. Oper. Res. **9** (1984), 441–458.
- [32] M.I. Reiman, *Some diffusion approximations with state space collapse*, Modelling and Performance Evaluation Methodology (F. Baccelli and G. Fayolle, eds.), Springer-Verlag, 1984, pp. 209–240.
- [33] D. Saure, P. Glynn, and A. Zeevi, *A linear programming-based algorithm for computing the stationary distribution of semi-martingale reflected Brownian motion*, Working paper, Columbia University, New York, NY, 2008.
- [34] A. L. Stolyar, *On the stability of multiclass queueing networks: A relaxed sufficient condition via limiting fluid processes*, Markov Processes and Related Fields **1** (1995), 491–512.
- [35] T. Tezcan, *Optimal control of distributed parallel server systems under the Halfin and Whitt regime*, Mathematics of operations research **33** (2008), no. 1, 51–90.
- [36] W. Whitt, *Weak convergence theorems for priority queues: Preemptive-resume discipline*, Journal of Appl. Prob. **8** (1971), 74–94.
- [37] _____, *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*, Springer-Verlag, New York., 2002.
- [38] R. J. Williams, *Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse*, Queueing Systems **30** (1998), 27–88.
- [39] _____, *An invariance principle for semimartingale reflecting Brownian motion*, Queueing Systems **30** (1998), 5–25.
- [40] H.Q. Ye and D.D. Yao, *Diffusion limit of a two-class network: stationary distributions and interchange of limits*, ACM SIGMETRICS Performance Evaluation Review **38** (2010), no. 2, 18–20.