The Annals of Applied Probability
2018, Vol. 28, No. 6, 3652–3697
https://doi.org/10.1214/18-AAP1401
© Institute of Mathematical Statistics, 2018

JUSTIFYING DIFFUSION APPROXIMATIONS FOR MULTICLASS QUEUEING NETWORKS UNDER A MOMENT CONDITION

BY HENG-QING YE[1] AND DAVID D. YAO[2]

Hong Kong Polytechnic University and Columbia University

Multiclass queueing networks (MQN) are, in general, difficult objects to study analytically. The diffusion approximation refers to using the stationary distribution of the diffusion limit as an approximation of the diffusion-scaled process (say, the workload) in the original MQN. To validate such an approximation amounts to justifying the interchange of two limits, $t \to \infty$ and $k \to \infty$, with t being the time index and k, the scaling parameter. Here, we show this interchange of limits is justified under a p^*th moment condition on the primitive data, the interarrival and service times; and we provide an explicit characterization of the required order (p^*), which depends naturally on the desired order of moment of the workload process.

1. Introduction. Multiclass queueing networks (MQN) have in recent decades become a popular model to study a wide range of engineering and service systems. The dynamics in such networks are typically driven by discrete events which occur in a stochastic manner: arrivals of jobs or customers, their service completions at one node (resource) followed by transitions to the next node in the network, occasional or sudden resource breakdowns, and so forth. To evaluate the steady-state performance of such networks, simulation often appears to be the only viable approach; whereas analytical methods mostly apply to stylized models with limited features.

To overcome this handicap, so-called diffusion approximations (or heavy-traffic steady-state approximation) have gained prominence in research and applications alike. The idea is to scale, in both time and space, the stochastic processes of interest (e.g., those associated with queue lengths and workloads) in the original network, In many cases, it can be shown, under the so-called *heavy traffic* condition (meaning the traffic intensity is such that resources are heavily utilized approaching capacity saturation), the scaled processes approach certain limiting regimes characterized by diffusion processes. The latter are more accessible analytically or computationally (e.g., [17]), and can thus serve as useful approximations for the original processes.

Received August 2015; revised January 2018.

[1] Supported in part by HK/RGC Grant 15508114.

[2] Supported in part by NSF Grant CMMI-1462495.

MSC2010 subject classifications. Primary 60K25, 90B15; secondary 60F17, 60J20.

Key words and phrases. Multiclass queueing network, diffusion limit, interchange of limits, uniform stability.

$$\begin{tikzcd}
\hat{W}^k(t) \arrow[r, "\text{II: } t \to \infty"] \arrow[d, "\text{I: } k \to \infty"'] & \hat{W}^k(\infty) \arrow[d, "\text{IV: } k \to \infty"] \\
\hat{W}(t) \arrow[r, "\text{III: } t \to \infty"'] & \hat{W}(\infty)
\end{tikzcd}$$

FIG. 1. *Interchange of limits.*

To illustrate this idea more formally, and following the setup due originally to Gamarnik and Zeevi [25], we refer to the rectangle depicted in Figure 1. Let $W(t)$, a vector process, denote the workload at time t in the original network. For technical and conceptual reasons, we consider an infinite sequence of copies or variations of the original network, indexed by k. Hence, let $W^k(t)$ denote the workload associated with the kth network in the sequence; and let $\hat{W}^k(t) := W^k(k^2t)/k$ denote its diffusion-scaled version. Establishing the diffusion limit, $\hat{W}(t) := \lim_{k\to\infty} \hat{W}^k(t)$, under heavy traffic is the task designated to the left vertical side, edge I, of the rectangle. Next, for each k, we want to establish the claim that $\hat{W}^k(t)$ has a stationary distribution as $t \to \infty$, and let $\hat{W}^k(\infty)$ denote the random variable associated with this limiting distribution. This step is represented on edge II of the rectangle. Completely analogous and represented on edge III of the rectangle is the claim that the diffusion limit $\hat{W}(t)$ also has a stationary distribution, embodied by $\hat{W}(\infty)$. The diffusion approximation is then to use this last stationary distribution, that of the diffusion limit $\hat{W}(t)$, as an approximation for the stationary distribution of the workload in the original network. This is tantamount to claiming that $\hat{W}^k(\infty)$ converges weakly, as $k \to \infty$, to $\hat{W}(\infty)$, which is represented on edge IV of the rectangle.

In other words, the diffusion approximation is a claim that starting from the workload process of the original network (kth in the sequence), going through edges I and III will reach the same result as going through edges II and IV. Formally, this can be expressed as

$$\lim_{t\to\infty} \lim_{k\to\infty} \hat{W}^k(t) = \lim_{k\to\infty} \lim_{t\to\infty} \hat{W}^k(t). \tag{1.1}$$

Thus, to justify the diffusion approximation boils down to justifying the interchange of two limits, $k \to \infty$ and $t \to \infty$.

Viewed this way, the four edges of the rectangle in Figure 1 constitute the key ingredients in the diffusion approximation of a stochastic network. First, we need to establish the diffusion limit (edge I); next, we need to make sure both the original network and the diffusion limit have stationary distributions (edges II, III); finally, to close the loop, we need to justify the interchange of the limits (edge IV).

1.1. *Contributions and organization.* The primary goal of this paper is to show that under certain moment conditions on the primitive processes (i.e., those concerning arrival and service mechanisms) in the original network, the interchange of the limits is justified for broad classes of networks. Specifically, we shall focus on a general class of MQN, for which the diffusion limit has been a well-studied subject (e.g., [4, 6, 12–14]).

Another equally important contribution of this paper is to bring forth a systematic approach—indeed a "recipe"—for the justification of diffusion approximations, developed here and in a companion paper [43]. To start with, suppose the diffusion limit exists, that is, edge I has already been established (otherwise, there will be nothing to "justify"). A powerful tool for establishing such a limit for various stochastic processing networks, the *hydrodynamics* approach, is given in [4, 39]. Then edges II and III can be established simultaneously together as follows. Consider the deterministic counterparts of the prelimit process $\hat{W}^k(t)$ and its diffusion limit $\hat{W}(t)$, denoted $\hat{w}^k(t)$ and $\hat{w}(t)$, respectively. [These are obtained by replacing the free processes in $\hat{W}^k(t)$ and $\hat{W}(t)$ by their drifts.] Then, following Dupuis and Williams [21], in order to show that the diffusion limit has a stationary distribution (edge III), it suffices to show the stability of $\hat{w}(t)$. Furthermore, we have recently established in [43] that the stability of $\hat{w}(t)$ leads to *uniform stability*; meaning, starting with a total initial workload that is bounded (by a single unit, say), the workload $\hat{w}^k(t)$ associated with the kth network, for any sufficiently large k, can be drained by a time that is independent of k. Then, by invoking the results in [19], this uniform stability leads to the stability of $\hat{W}^k(t)$; and hence, edge II.

Next, to establish edge IV, the crucial step is to bound the *pth moment of the workload process* ($p > m + 1$, when the convergence of the mth moment of the workload process is required). To do so, in [43], we have introduced a *bounded workload condition* (more on this below); whereas in this paper, we will show this can be accomplished by requiring the *p^*th moment condition*, a moment condition of suitable order on the primitive processes [specifically, the order is $p^* > 2(p + 2)$]. Once this is done, along with the uniform stability established on edge II, we can prove the uniform pth moment stability of the workload processes (refer to Lemma 3.9 below), which will lead to the tightness of $\{\hat{W}^k(\infty)\}$ and the convergence (of stationary distributions or moments) on edge IV, following the approach (by now standard) in [9, 19, 43].

Indeed, the above recipe can be readily applied to many other stochastic network models as well, such as those in [11, 18, 28, 32, 35, 42]. In all those models, the networks can be represented by strong Markov processes, for which standard theories (e.g., ergodicity) are available, and diffusion limits have already been established in the form of (semimartingale) reflected Brownian motion. The justification of diffusion approximations can then be carried out by following the recipe outlined above.

Below, we start with a review of related literature in the remaining part of this section. We then present in Section 2 the class of MQN that we shall focus on, along with results regarding edges I, II and III in Figure 1 (in Section 2.1). The results concerning the interchange of limits are presented in Section 2.2, and the detailed proofs are collected in Section 3. To facilitate exposition, secondary results and their proofs are collected in Appendices A and B.

1.2. *Related literature.* A brief review of the related literature is in order. Recent studies on the interchange of limits to justify the diffusion approximation have been initiated by Gamarnik and Zeevi [25] and Budhiraja and Lee [9], where they have established the justification for the generalized Jackson network, a single-class queueing network. Their results and approaches are then refined for some multiclass networks in [26, 29, 30, 34, 43]. Other related works involve the many-server regime, including [7, 8, 16, 23, 24, 36–38]. This is a different limiting regime and often calls for a very different approach (from those used in MQN), such as Stein's method in [7, 8]. Common to all studies, however, is to identify key conditions of the systems under study so as to establish certain properties (e.g., the moment bound of the workloads, as mentioned above) that directly lead to the interchange of limits. Yet, such conditions (e.g., the Lipschitz continuity of Skorohod mapping, the explicit bound on the workload in terms of interarrival and service times, etc.) are often not present in multiclass networks; and this has become a major handicap in applying diffusion approximation to a wider range of stochastic processing networks. To extend such conditions, or more precisely, to identify new conditions that will lead to the justification of the interchange of limits in MQN, has become the focus of recent works including [26, 43] as well as our current study.

In the multiclass setting, most related to our work is the recent study by Gurvich [26], which aims to develop a systematic approach to the interchange of limits for a class of multiclass queueing networks operating under the queue-ratio service discipline, as well as the more standard buffer-priority discipline (see Example 2.2 in [26], which also refers to [12, 14]). The key step developed there (to justify the interchange) is to verify a condition that the fluid model associated with the sequence of networks under heavy traffic converges to the fixed-point state space (also known as the "invariant manifold") at a *linear rate*. This condition is stronger than the usual uniform attraction property of the same fluid model. In [26], finite moments of order p, for all values of $p > 0$, are assumed on the primitives, with the following remark after presenting its main result, Theorem 3.1: "However, the mapping from the value of p … to the number of moments, m, for which the convergence holds is not as clear as in the generalized Jackson case . . . , where . . . such convergence holds for all $m < p - 1$."

Here, we show that a moment condition on the primitives is all that's needed to justify the interchange, given the existence of the diffusion limit and its stationary distribution. Thus, in particular, the linear-rate convergence condition in [26] for networks under the queue-ratio discipline can be removed. Furthermore, we provide an explicit characterization of the relation between p and m mentioned in the above quote from [26]. The required order of moments on the primitives is $p^* > 2(p + 2)$, where p relates to m via $m < p - 1$ as in the above quote from [26]. In our approach, however, p also specifically refers to the pth moment of the workload that we want to bound, a key step leading to the interchange as motivated earlier. In other words, we use this p as a bridge to connect the desired order of convergence m with the required order of moments p^* on the primitives. For more details, refer to the p^*th moment condition below in (2.38).

Similarly, in our own recent work [43], we bound the pth moment of workload processes with a condition that the workload can be bounded by a "free process" plus the initial workload. This bounded workload condition does not imply the p^*th moment condition developed here, and is not implied by the latter either. However, to verify the bounded workload condition requires effort as illustrated via several examples in [43]. (Once the bounded workload condition is verified, however, the interchange is justified without requiring any higher order moments on the primitives.) In contrast, the p^*th moment condition here is trivial to verify, and indeed automatically holds in networks where the primitives have moments of all orders (e.g., renewal arrivals with phase-type interarrival times and i.i.d. phase-type service times).

Using the p^*th moment condition to justify the interchange of limits, however, turns out to be a serious challenge. To prove the boundedness of the pth moment of workload processes, we need to focus on a sequence of "regular" events, under which the network processes behave "nicely," and the probabilities of these events occurring approach 1 at a certain rate (cf. Lemma 3.1). We then apply Bramson's "hydrodynamic" approach (e.g., [4, 26, 32, 42]) to show that the bounded workload condition holds for sample paths in the regular events (cf. Lemma 3.5). Therefore, the workload processes, when restricted to the regular events, possess a bounded pth moments. On the other hand, the pth moment of the workload processes restricted to the small-probability, nonregular events, have the same bound. Combining these two cases leads to the desired result—to justify the interchange of limits via bounding the pth moment of workload processes.

2. Multiclass queueing networks. Let $\mathcal{L} = \{1, \dots, L\}$ denote the set of servers, and $\mathcal{R} = \{1, \dots, R\}$ the set of job classes. Jobs move from one server (resource) to another sequentially, and every job will have a distinct class index at different servers. This labeling of job classes can be conveniently captured by a "constituent" (incidence) matrix $C = (C_{\ell r})_{\ell \in \mathcal{L}, r \in \mathcal{R}}$, where $C_{\ell r} = 1$ if class-r

jobs are served by server ℓ, and $C_{\ell r} = 0$ otherwise. Note, each column of C has one and only one entry that equals 1, since each job class is served by one server and one server only; and every job's class identity (r) will uniquely determine its server (ℓ, such that $C_{\ell r} = 1$). Whenever a server is available, the service discipline will decide which job (if any) to be processed from the classes associated with the server. We will focus on the head-of-line (HOL) service discipline; examples include preemptive resume static buffer priority (SBP, [12, 14]), first-in-first-out [4, 13], head-of-line proportional processor sharing [4] and the queue-ratio discipline [26], among others. After service completion, a class-r job turns into a class-s job with probability p_{rs}, or leaves the network with probability $1 - \sum_s p_{rs}$. Assume this class transition scheme to be independent among all job classes and independent of all other arrival and service mechanisms. Denote $P := (p_{rs})_{r,s\in\mathcal{R}}$. Clearly, P is a substochastic matrix; and we will assume it has a spectral radius < 1, and is thus invertible. (This means the network under study is an *open* network.)

We consider a sequence of networks, indexed by k. For the kth network, denote for each class r the interarrival times between consecutive and external jobs as $u_r^k(i)$, and denote the amount of work (service time) each job brings to the network as $v_r^k(i)$, $i = 1, 2, \ldots.$ Assume the interarrival and service times possess finite pth moments, $p > 2$. In particular, since we need to deal with systems that do not necessarily start empty, we reserve $u_r^k(1)$ and $v_r^k(1)$ to denote, at time zero, the *residual* time and work until the next arrival and the next service completion, respectively. Furthermore, we assume that $\{(u_r^k(i), v_r^k(i)), i \geq 2\}$ are i.i.d. with mean $((\alpha_r^k)^{-1}, \mu_r^{-1})$ and variance $((\sigma_{a,r}^k)^2, \sigma_{b,r}^2)$. Denote $\alpha^k = (\alpha_r^k)_{r\in\mathcal{R}}$ and $M = \operatorname{diag}(\mu_r^{-1})_{r\in\mathcal{R}}$. For ease of exposition, we have assumed all the networks share the same mean and variance of service times and hence omitted the index k from μ_r, $\sigma_{b,r}$ and M.

The routing of jobs is defined by a routing sequence $\{\phi_{rs}^k(i), i = 1, 2, \ldots\}$, where $\phi_{rs}^k(i) = 1$ if the ith class-r job is routed to class-s upon its service completion. Hence, we have

$$\mathsf{P}\{\phi_{rs}^k(i) = 1\} = p_{rs} \quad \text{and} \quad \mathsf{P}\{\phi_{rs}^k(i) = 0\} = 1 - p_{rs}.$$

Assume $\{\phi_{rs}^k(i)\}$ are independent random variables, and are independent of $\{(u_r^k(i), v_r^k(i))\}$, too.

Let $\lambda^k = (I - P^T)^{-1}\alpha^k$, which is the solution to the traffic equation, $\lambda^k = \alpha^k + P^T\lambda^k$. Let

$$\rho^k := (\rho_\ell^k)_{\ell\in\mathcal{L}} = CM\lambda^k \bigl(= CM(I - P^T)^{-1}\alpha^k\bigr),$$

where ρ_ℓ^k is the traffic intensity for server ℓ, $\ell \in \mathcal{L}$.

The three primitive processes that drive the kth network are the *delayed* (i.e., including the residuals) renewal processes associated with the external job arrivals and the service times of the jobs and the routing process. Specifically, let $E^k(t) =$

$(E_r^k(t))_{r\in\mathcal{R}}$ and $S^k(t)=(S_r^k(t))_{r\in\mathcal{R}}$, $t\ge 0$, where

$$E_r^k(t)=\max\left\{i:\sum_{j=1}^{i}u_r^k(j)\le t\right\}\quad\text{and}$$
$$S_r^k(t)=\max\left\{i:\sum_{j=1}^{i}v_r^k(j)\le t\right\}. \tag{2.1}$$

The routing process, denoted $\Phi_r^k(n):=(\Phi_{rs}^k(n))_{s\in\mathcal{R}}$ for class-r, is defined by the routing sequence as

$$\Phi_{rs}^k(n)=\sum_{i=1}^{n}\phi_{rs}^k(i),$$

which counts the number of jobs, among the first n jobs served by serve r, routed to server s.

With the residuals $(u_r^k(1),v_r^k(1))_{r\in\mathcal{R}}$ removed, the (undelayed) arrival and service processes are denoted: $E^{o,k}(t)=(E_r^{o,k}(t))_{r\in\mathcal{R}}$ and $S^{o,k}(t)=(S_r^{o,k}(t))_{r\in\mathcal{R}}$, $t\ge 0$, where

$$E_r^{o,k}(t)=\max\left\{i:\sum_{j=2}^{i}u_r^k(j)\le t\right\}\quad\text{and}$$
$$S_r^{o,k}(t)=\max\left\{i:\sum_{j=2}^{i}v_r^k(j)\le t\right\}. \tag{2.2}$$

Here and below, the superscript "o" denotes the undelayed version of a (possibly) delayed renewal process. By default, assume $E_r^k(t)\equiv 0$ [or $u_r^k(1)=\infty$] if class r has no external arrivals (i.e., $\alpha_r^k=0$).

The main performance measures of interest are the queue length process $Q^k(t)=\{Q_r^k(t)\}_{r\in\mathcal{R}}$, and the workload process $W^k(t)=\{W_\ell^k(t)\}_{\ell\in\mathcal{L}}$. Specifically, $Q_r^k(t)$ counts the number of class-r jobs in the (kth) network at time t, and $W_\ell^k(t)$ translates the job count into the amount of work server ℓ is facing. Let $D_r^k(t)$ be the total amount of service time allocated to serving class-r jobs during $(0,t]$, and call $D^k(t)=\{D_r^k(t)\}_{r\in\mathcal{R}}$ the *allocation process.*

The dynamics of the (kth) network can be represented as follows:

$$Q^k(t)=Q^k(0)+E^k(t)+\sum_{r=1}^{R}\Phi_r^k(S_r^k(D_r^k(t)))-S^k(D^k(t))\ge 0; \tag{2.3}$$

(2.4) $D^k(t)$ is nondecreasing, and
satisfies conditions specific to the service discipline;

$$W^k(t)=CMQ^k(t)\ge 0; \tag{2.5}$$

$$Y^k(t) = et - CD^k(t) \text{ is nondecreasing, with } Y^k(0) = 0; \tag{2.6}$$

$$\int_0^\infty W^k(t)\, dY^k(t) = 0. \tag{2.7}$$

Here, the composition $S^k(D^k(t))$ is an R-dimensional process with the rth component being $S_r^k(D_r^k(t))$. Note that $S^k(D^k(t)) = \{S_r^k(D_r^k(t))\}_{r\in\mathcal{R}}$ and $S_r^k(D_r^k(t))$ counts the total number service completions for class-r jobs by time t, $Y_\ell^k(t)$ keeps track of the cumulative idle time of server ℓ during the time interval $[0, t]$, and the last equality is the work-conserving (or nonidling) condition.

As an example of the specification in (2.4), the static buffer priority discipline can be characterized as follows (cf. [12, 14]). For each class r, denote H_r as the set of job classes that are served by the same server associated with r and have a priority no less than r; and let $\varsigma_r^k(t) = 1$ if $Q_r^k(t) > 0$ and $\sum_{r'\in H_r\setminus\{r\}} Q_{r'}^k(t) = 0$, and $\varsigma_r^k(t) = 0$ otherwise. That is, the function $\varsigma_r^k(t)$ indicates whether the class r is being served by its associated server at time t. Then the allocation can be written as

$$D_r^k(t) = \int_0^t \varsigma_r^k(s)\, ds, \qquad t \geq 0. \tag{2.8}$$

We follow the standard approach (e.g., [9, 15, 25, 26]) to construct a Markov process representation of the network, denoted $\Xi^k(t)$, by appending to the queue-length process other state information such as the residual interarrival and service times (or the "age" of each existing job in the FIFO case). Under a HOL discipline, $\Xi^k(t)$ is piecewise-deterministic and satisfies the strong Markov property [20], which is essential for the stability of the fluid model to imply the stability (positive Harris recurrence) of the original network [15, 33].

Take the static buffer priority discipline as an example again to elaborate the definition of $\Xi^k(t)$. Denote $U^k(t) = (U_r^k(t))_{r\in\mathcal{R}}$ and $V^k(t) = (V_r^k(t))_{r\in\mathcal{R}}$, $t \geq 0$, where

$$U_r^k(t) = \sum_{i=1}^{E_r^k(t)+1} u_r^k(i) - t, \qquad V_r^k(t) = \sum_{i=1}^{S_r^k(D_r^k(t))+1} v_r^k(i) - D_r^k(t). \tag{2.9}$$

That is, at any given time t, for class r, $U_r^k(t)$ is the remaining time before the next external arrival, and $V_r^k(t)$ is the remaining service time for the job that is in service. (If there is no class r job at the time, $V_r^k(t)$ is the service time for the arriving class r job.) Note, at time $t = 0$, we have $U_r^k(0) = u_r^k(1)$ and $V_r^k(0) = v_r^k(1)$, the residuals at time zero introduced above. Hence, below we shall refer to $U_r^k(t)$ and $V_r^k(t)$ as "residuals" (at t) as well. Then $\Xi^k(t) = (Q^k(t), U^k(t), V^k(t))$ is a strong Markov process, taking values on the nonnegative orthant of the $3R$-dimensional Euclidean space, denoted $\mathcal{X}$ (cf. [15, 20, 29]). Clearly, the dynamics of the Markov process $\Xi^k(t)$ will be completely determined when the initial state is given. Below, we will often consider many copies of the same network, each starting from a different initial state. To highlight the dependence on the initial state, we will append it to the argument of the corresponding Markov process and workload process. Hence, instead of $\Xi^k(t)$ [resp. $Q^k(t)$, $W^k(t)$, $D^k(t)$ and $Y^k(t)$], wherever necessary we will write $\Xi^k(t; x)$ [resp. $Q^k(t; x)$, $W^k(t; x)$, $D^k(t; x)$ and $Y^k(t; x)$] with $x = \Xi^k(0) \in \mathcal{X}$ being the initial state.

2.1. *Diffusion limit.* We assume that as $k \to \infty$, the parameters α_r^k, λ_r^k, $\sigma_{a,r}^k$ and ρ_r^k converge to α_r, λ_r, $\sigma_{a,r}$ and ρ_r, respectively; and thus these parameters satisfy $\rho = CM\lambda = CM(I - P^T)^{-1}\alpha$. The *heavy traffic condition* is in force when $\rho = e$ and

$$k(\rho^k - e) \to \gamma \qquad \text{as } k \to \infty, \tag{2.10}$$

for some L-dimensional constant vector γ. That is, the total traffic load on each server, including both external arrivals and internal transfers, is equal to its service capacity (asymptotically).

Recall, we require the primitives of the network, the interarrival and service times, to possess a finite pth moment. Now for a *sequence* of networks, we need to strengthen this condition so that it holds *uniformly* for all the networks. To avoid technicalities, we assume that the network sequence is driven by the same primitives except the initial arrival and service times; that is, assume for all k,

$$\alpha_r^k u_r^k(i) = \alpha_r^1 u_r^1(i) \quad \text{and} \quad v_r^k(i) = v_r^1(i), \qquad i \geq 2, r \in \mathcal{R}. \tag{2.11}$$

For the given $p > 2$, assume all interarrival and service times have bounded pth moments:

$$\mathsf{E} \sum_{r \in \mathcal{R}} [(u_r^1(2))^p + (v_r^1(2))^p] < \infty. \tag{2.12}$$

In addition, for each k and $r \in \mathcal{R}$, we assume that

$$\mathsf{P}\{u_r^k(2) \geq a\} > 0 \qquad \text{for any } a > 0; \tag{2.13}$$

and that for some integer $j \geq 2$ and some nonnegative function $p(x)$ satisfying $\int_0^\infty p(s) > 0$, the following inequality holds:

$$\mathsf{P}\left\{a \leq \sum_{i=2}^{j} u_r^k(i) \leq b\right\} \geq \int_a^b p(x)\,dx \qquad \text{for any } 0 \leq a < b. \tag{2.14}$$

The above are certain forms of a “spread-out” condition, required to guarantee the positive (Harris) recurrence and hence the uniqueness of the stationary distribution of the prelimit networks in edge II of Figure 1. They also appeared in prior works, for example, [5, 15].

Apply the standard diffusion scaling (along with centering) to the primitive and derived processes:

$$
\begin{aligned}
(\hat{E}_r^{o,k}(t), \hat{S}_r^{o,k}(t)) &= \frac{1}{k}(E_r^{o,k}(k^2t) - \lambda_r^k k^2 t, S_r^{o,k}(k^2t) - (\nu_r^k)^{-1}k^2t),\\
(\hat{E}_r^{k}(t), \hat{S}_r^{k}(t)) &= \frac{1}{k}(E_r^{k}(k^2t) - \lambda_r^k k^2 t, S_r^{k}(k^2t) - (\nu_r^k)^{-1}k^2t),\\
\hat{\Phi}_{rs}^k(t) &= \frac{1}{k}(\Phi_{rs}^k(\lfloor k^2t \rfloor) - p_{rs}k^2t),\\
(\hat{\Xi}_r^k(t), \hat{Q}_r^k(t), \hat{W}_r^k(t)) &= \frac{1}{k}(\Xi_r^k(k^2t), Q_r^k(k^2t), W_r^k(k^2t)).
\end{aligned}
\tag{2.15}
$$

We will use the following so-called fluid scaling of the primitive processes as well:

$$
\begin{aligned}
&(\bar{E}_r^{o,k}(t), \bar{S}_r^{o,k}(t), \bar{\Phi}_{rs}^k(t), \bar{E}_r^k(t), \bar{S}_r^k(t))\\
&\quad = \frac{1}{k}(E_r^{o,k}(kt), S_r^{o,k}(kt), \Phi_{rs}^k(\lfloor kt \rfloor), E_r^k(kt), S_r^k(kt)).
\end{aligned}
\tag{2.16}
$$

We now apply diffusion scaling to the equations in (2.3)–(2.7) and the related processes. Particularly, the equation in (2.3) becomes

$$
\hat{Q}^k(t) = \hat{Q}^k(0) + \hat{\xi}^k(t) + k(\alpha^k - \alpha)t + k[\alpha t - (I - P^T)M^{-1}\tilde{D}^k(t)],
\tag{2.17}
$$

where

$$
\begin{aligned}
\hat{\xi}^k(t) &= \hat{E}^k(t) + \sum_{r=1}^R \hat{\Phi}_r^k(\tilde{S}_r^k(\tilde{D}_r^k(t))) - (I - P^T)\hat{S}^k(\tilde{D}^k(t)),\\
\tilde{S}^k(t) &= \frac{1}{k^2}S^k(k^2t), \qquad \tilde{D}^k(t) = \frac{1}{k^2}D^k(k^2t),\\
\hat{S}^k(\tilde{D}^k(t)) &= \{\hat{S}_r^k(\tilde{D}_r^k(t))\}_{r\in\mathcal{R}}.
\end{aligned}
$$

Associated with a given service discipline, there exists a so-called lifting matrix, denoted Δ_q, which is a $(R \times L)$-matrix that maps the workload to the queue length in the diffusion limit. This matrix is specified through the uniform attraction property, which will be discussed shortly. For example, consider the static buffer priority discipline: $(\Delta_q)_{\ell r} = \mu_r$, if class r has the lowest priority among all classes served by server ℓ; and $(\Delta_q)_{\ell r} = 0$, otherwise. Denote $\hat{\varepsilon}^k(t) = \hat{Q}^k(t) - \Delta_q \hat{W}^k(t)$. Then, following [39] (Section 4), we can reduce the dynamics of the workloads under diffusion scaling to the following:

$$
\hat{W}^k(t) = \Psi CM(I - P^T)^{-1}(\hat{Q}^k(0) - \hat{\varepsilon}^k(t)) + \hat{X}^k(t) + \Psi\hat{Y}^k(t),
\tag{2.18}
$$

(2.19) $\hat{Y}^k(t)$ is nondecreasing, with $\hat{Y}^k(0) = 0$,

$$
\int_0^\infty \hat{W}^k(t)\, d\hat{Y}^k(t) = 0,
\tag{2.20}
$$

where the reflection matrix is given as $\Psi = [CM(I-P^T)^{-1}\Delta_q]^{-1}$ and the free process as

$$\hat{X}^k(t) = \Psi CM\big(I-P^T\big)^{-1}\hat{\xi}^k(t) + k\Psi\big(\rho^k - e\big)t. \tag{2.21}$$

To understand the equation in (2.18), we first write $\hat{Q}^k(t) = \Delta_q \hat{W}^k(t) + \hat{\varepsilon}^k(t)$ in (2.17), and then multiply both sides with $\Psi CM(I-P^T)^{-1}$ to get the following:

$$\begin{aligned}\hat{W}^k(t) &= \Psi CM\big(I-P^T\big)^{-1}\big(\hat{Q}^k(0) - \hat{\varepsilon}^k(t)\big) + \Psi CM\big(I-P^T\big)^{-1}\hat{\xi}^k(t)\\ &\quad + k\Psi CM\big(I-P^T\big)^{-1}\big(\alpha^k - \alpha\big)t + k\Psi\big[CM\big(I-P^T\big)^{-1}\alpha t - C\tilde{D}^k(t)\big].\end{aligned}$$

Taking into account that $CM(I-P^T)^{-1}\alpha^k = \rho^k$ and $CM(I-P^T)^{-1}\alpha = e$, and that $\hat{Y}^k(t) = k[et - C\tilde{D}^k(t)]$ [cf. (2.6)], we know the above equation can be written as the one in (2.18) immediately.

Here, the existence of the reflection matrix Ψ is implicitly assumed. Also assumed is Ψ being a complete-S matrix (cf. [39]), which is necessary for the existence of the diffusion limit in the theorem below.

In addition, the following fluid model associated with the original network can be derived using the hydrodynamic approach (refer to Section 3 for a more general version):

$$\bar{q}(t) = \bar{q}(0) + \alpha t - \big(I-P^T\big)M^{-1}\bar{d}(t), \tag{2.22}$$

$$\bar{w}(t) = CM\bar{q}(t) \geq 0, \tag{2.23}$$

$$\bar{y}(t) = et - CM\bar{d}(t), \qquad \bar{d}(t) \text{ and } \bar{y}(t) \text{ are nondecreasing}, \tag{2.24}$$

$$\int_0^\infty \bar{w}(t)\,d\bar{y}(t) = 0, \tag{2.25}$$

(2.26) additional conditions specifying the service discipline.

To illustrate the specification in (2.26), continue with the static buffer priority discipline in (2.8), with the allocation process $\bar{d}(t) = \{\bar{d}_r(t)\}_{r\in\mathcal{R}}$ expressed as

$$\bar{d}_r(t) = \int_0^t \bar{\varsigma}_r(s)\,ds, \qquad t \geq 0, \tag{2.27}$$

where $\{\bar{\varsigma}_r(t)\}_{r\in\mathcal{R}}$ are nonnegative functions satisfying the following for $t \geq 0$:

$$\sum_{r\in\mathcal{R}} \bar{\varsigma}_r(t) \leq 1 \quad \text{and} \quad \sum_{r\in H_r} \bar{\varsigma}_r(t) = 1 \qquad \text{if } \bar{q}_r(t) > 0.$$

Refer to [4, 12, 14] for more details regarding the static buffer priority discipline (and other HOL disciplines as well) in the fluid model.

We say that the uniform attraction property holds with the lifting matrix Δ_q if there exists a positive real function $h(t)$, satisfying $\lim_{t\to\infty} h(t) = 0$, such that for any solution to the above equations with $|q(0)| \leq 1$, the following holds:

$$\big|\bar{q}(t) - \Delta_q w(t)\big| \leq h(t) \qquad \text{for all } t \geq 0. \tag{2.28}$$

An alternative (and equivalent) statement for the above condition is for some $w^* \geq 0$,

$$|\bar{q}(t) - \Delta_q w^*| \leq h(t) \qquad \text{for all } t \geq 0. \tag{2.29}$$

The uniform attraction property has been established for multiclass queueing networks under various service disciplines; refer to [2–4, 6, 12–14]; details of the associated lifting matrices can be found in these papers, too.

In the following theorem, we will study the weak convergence (denoted as "⇒") of the diffusion-scaled processes in the Skorohod space, the space of RCLL functions. Strictly speaking, we need to deal with the Skorohod metric [1, 39]. However, since all the limiting processes involved are continuous processes (Brownian motions), and the u.o.c. convergence to a continuous function implies the convergence under the Skorohod metric [1], it is convenient (and indeed equivalent in the current context) to treat the Skorohod space as endowed with the more familiar uniform metric.

THEOREM 2.1 (Diffusion limit [4, 39]). *Suppose the heavy traffic condition in* (2.10) *is in force*; *the initial workload and the initial residuals satisfy the following*:

$$(\hat{W}^k(0), \hat{Q}^k(0)) \quad \Rightarrow \quad (\hat{W}(0), \hat{Q}(0)) \qquad \text{with } \hat{Q}(0) = \Delta_q \hat{W}(0),$$

$$|\hat{U}^k(0)| + |\hat{V}^k(0)| = \frac{1}{k}(|u^k(1)| + |v^k(1)|) \to 0,$$

for some $\hat{W}(0)$; *and the uniform attraction property in* (2.28), (2.29) *holds*. *Then we have the following weak convergence when* $k \to \infty$:

$$(\hat{W}^k(t), \hat{Q}^k(t), \hat{X}^k(t), \hat{Y}^k(t)) \quad \Rightarrow \quad (\hat{W}(t), \hat{Q}(t), \hat{X}(t), \hat{Y}(t)),$$

with the limit characterized by the following equations:

$$\hat{W}(t) = \hat{W}(0) + \hat{X}(t) + \Psi \hat{Y}(t) \geq 0; \tag{2.30}$$

$$\hat{Y}_\ell(t) \text{ is nondecreasing in } t \geq 0, \qquad \hat{Y}_\ell(0) = 0, \qquad \ell \in \mathcal{L}; \tag{2.31}$$

$$\int_0^\infty \hat{W}_\ell(t)\, d\hat{Y}_\ell(t) = 0, \qquad \ell \in \mathcal{L}; \tag{2.32}$$

$$\hat{Q}(t) = \Delta_q \hat{W}(t). \tag{2.33}$$

Here, $\hat{X}(t)$ *is a Brownian motion with drift* $\theta := \Psi\gamma$ (*and properly defined covariance matrix*), *and* $\hat{W}(t)$ *is an semimartingale reflected Brownian motion* (*refer to* [39]).

Recall we have assumed in the above that the reflection matrix Ψ is well defined (i.e., the matrix $CM(I - P^T)^{-1}\Delta_q$ is invertible) and is a complete-S matrix, which guarantees the existence and uniqueness of the limit $\hat{W}(t)$; refer to [39]. We remark that to guarantee the convergence in the above theorem, instead of the condition in (2.12), it suffices to assume a weaker condition, Bramson's uniform second moment condition, on the primitive processes; refer to [4].

The following *dynamic complementarity problem* (DCP) is a deterministic version of the above limit [particular the first three equations in (2.30)–(2.32)], which is obtained by replacing $\hat{X}(t)$ in (2.30) with its drift term,

$$\hat{w}(t) = \hat{w}(0) + \theta t + \Psi \hat{y}(t) \geq 0; \tag{2.34}$$

$$\hat{y}_\ell(t) \text{ is nondecreasing in } t \geq 0, \qquad \hat{y}_\ell(0) = 0, \qquad \ell \in \mathcal{L}; \tag{2.35}$$

$$\int_0^\infty \hat{w}_\ell(t)\, d\hat{y}_\ell(t) = 0, \qquad \ell \in \mathcal{L}. \tag{2.36}$$

We shall refer to the deterministic DCP in (2.34)–(2.36) as *stable*, if there exists a time T such that for any solution $(\hat{w}(t), \hat{y}(t))$ with $|\hat{w}(0)| \leq 1$, we have $\hat{w}(t) = 0$ for all $t \geq T$.

Then the convergence results along edges II and III in Figure 1 are readily established following similar studies in the literature; and we summarize these results in the following theorem.

THEOREM 2.2. *Suppose the heavy traffic condition in* (2.10) *holds, and the DCP in* (2.34)–(2.36) *is stable. Then we have the following results*:

(a) *The diffusion limit $\hat{W}(t)$ in Theorem 2.1 is positive recurrent and has a unique stationary distribution.*

(b) *For any sufficiently large k, $\hat{\Xi}^k(t) = (\hat{Q}^k(t), \hat{U}^k(t), \hat{V}^k(t))$ is positive recurrent and has a unique stationary distribution. Furthermore, if the pth moment condition in* (2.12) *holds, then for any $m \in [0, p-1)$ and for sufficiently large k, the stationary workload has a finite mth moment and*

$$\lim_{t\to\infty} \mathsf{E}\big|\hat{W}^k(t;x)\big|^m = \mathsf{E}\big|\hat{W}^k(\infty)\big|^m < \infty \qquad \textit{for any initial state } x, \tag{2.37}$$

where $\hat{W}^k(\infty)$ stands for a random variable (*vector*) *following the stationary distribution of $\hat{W}^k(t)$.*

Indeed, the conclusion in (a) is due to Dupuis and Williams [21]. The key to establishing the edge II result is the concept of *uniform stability*, which refers to the property that if the deterministic DCP corresponding to the diffusion limit is stable, the fluid models corresponding to the prelimit networks are also stable. (Also refer to Lemma 3.7 below.) Consequently, the conclusion in (b), follows from Dai and Meyn [19].

2.2. *Interchange of limits.* Here in this section, we establish edge IV (convergence of stationary distributions) in Figure 1. Furthermore, given $p > 2$, we also establish the convergence of the mth moment of the stationary workloads, with $0 < m < p - 1$. To this end, we need the following condition, which strengthens the pth moment condition of Theorem 2.2.

p^*th moment condition: All interarrival and service times have bounded p^*th moments, that is, for some $p^* > 2(p+2)$,

$$\mathsf{E}\sum_{r\in\mathcal{R}}\big[(u_r^1(2))^{p^*} + (v_r^1(2))^{p^*}\big] < \infty. \tag{2.38}$$

This condition will guarantee the boundedness of the pth moment of workload, which holds the key to proving the convergence of stationary mth moments of workload for $m < p - 1$.

Note that the above p^*th moment condition *implies* the following slightly weaker form: for some constant $\kappa > 0$ and for all $t \geq 0$,

$$\mathsf{E}\sup_{0\leq s\leq t}\sum_{r\in\mathcal{R}}\big(|E_r^{o,k}(s) - \lambda_r^k s|^{p^*} + |S_r^{o,k}(s) - \mu_r^k s|^{p^*}\big) \leq \kappa\big(1 + t^{p^*/2}\big), \tag{2.39}$$

and furthermore,

$$\mathsf{E}\sup_{0\leq s\leq t}\sum_{r\in\mathcal{R}}\big(|\hat{E}_r^{o,k}(s)|^{p^*} + |\hat{S}_r^{o,k}(s)|^{p^*}\big) \leq \kappa\big(1 + t^{p^*/2}\big). \tag{2.40}$$

The above variation is technically convenient and has been used in previous studies [9, 43]. To prove the claimed implication, refer to Appendix A.1, Lemma A.3.

THEOREM 2.3. *Suppose the heavy traffic condition in* (2.10) *and the uniform attraction property in* (2.28), (2.29) *hold. Assume furthermore the stability of the deterministic DCP in* (2.34)–(2.36) *and the p^*th moment condition. Then the following weak convergence holds*:

$$\big(\hat{W}^k(\infty), \hat{Q}^k(\infty)\big) \quad \Rightarrow \quad \big(\hat{W}(\infty), \hat{Q}(\infty)\big) \qquad \text{as } k \to \infty. \tag{2.41}$$

Furthermore, for any $m \in [0, p - 1)$, we have

$$\big(\mathsf{E}|\hat{W}^k(\infty)|^m, \mathsf{E}|\hat{Q}^k(\infty)|^m\big) \to \big(\mathsf{E}|\hat{W}(\infty)|^m, \mathsf{E}|\hat{Q}(\infty)|^m\big) \qquad \text{as } k \to \infty. \tag{2.42}$$

Note, the above theorem is in parallel to Theorem 14 of [43], which also justifies the interchange of limits but under a different condition, the so-called *bounded workload condition*: for some constant $\kappa > 0$,

$$\sup_{0\leq s\leq t}|\hat{W}^k(s)| \leq \kappa\Big(|\hat{W}^k(0)| + \sup_{0\leq s\leq t}|\hat{X}^k(s)|\Big). \tag{2.43}$$

In particular, it is shown in [43] the above condition implies the boundedness of the pth moment of the workload process,

$$\mathsf{E} \sup_{0\le s\le t} |\hat{W}^k(s)|^p \le \kappa\big(|\hat{\Xi}^k(0)|^p + 1 + t^p\big), \tag{2.44}$$

under the bounded pth moment condition of primitives in (2.12). This bound on the workload holds the key to establishing other properties for justifying the interchange of limits.

Clearly, the bounded workload condition in (2.43) and the p^*th moment condition in (2.38) do not imply each other: the latter is trivial to verify, being imposed directly on the *primitives*; whereas the former requires (slightly) lower moments on the primitives but could be difficult to verify.

3. Proof of Theorem 2.3.

We break the proof into three subsections. In Section 3.1, we identify certain *regular* events associated with "nice" sample paths, such as the fluid-scaled arrival processes lying within a certain range of their mean values; and develop bounds on the probabilities of these regular events. Then, in Section 3.2, we demonstrate that the workload process, too, behaves "nicely" under the regular events. To do so, we apply Bramson's "hydrodynamic" approach [4] to show that the bounds in (2.43) work for sample paths under the regular events. Such a bound in the regular event, combined with a crude bound for the pth moment of the workload under the "nonregular" (or, rare) events, leads to the pth moment bound of the workload in (2.44) under the diffusion scaling (Lemma 3.6). The rest of the proof in Section 3.3 is to establish standard properties of the workload process such as uniform integrability, uniform pth moment stability and tightness, which then complete the proof of Theorem 2.3.

A road map summarizing the above is illustrated in Figure 2. The part marked out by the dotted rectangle is the focus of this paper.

3.1. *Probability bound of regular events, complementarity, oscillation inequality, uniform continuity and uniform attraction.*

Probability bound of regular events. Define the variables:

$$u_r^{k,\max}(t) := \max\left\{u_r^k(i) : \sum_{i'=2}^{i-1} u_r^k(i') \le t, i = 2, 3, \dots\right\}, \tag{3.1}$$

$$v_r^{k,\max}(t) := \max\left\{v_r^k(i) : \sum_{i'=2}^{i-1} v_r^k(i') \le t, i = 2, 3, \dots\right\}. \tag{3.2}$$

The first variable is the maximal interarrival time of class r realized before time t for the kth network; the second variable is analogous, for the service times. Note that the initial residuals $u_r^k(1)$ and $v_r^k(1)$ are excluded. Let t^* and u^* be any positive

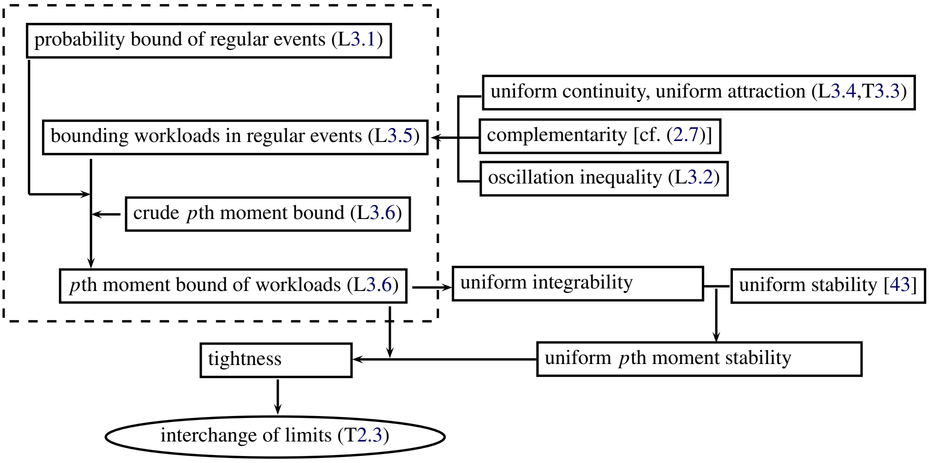
probability bound of regular events (L3.1)
uniform continuity, uniform attraction (L3.4,T3.3)
bounding workloads in regular events (L3.5)
complementarity [cf. (2.7)]
oscillation inequality (L3.2)
crude pth moment bound (L3.6)
pth moment bound of workloads (L3.6)
uniform integrability
uniform stability [43]
tightness
uniform pth moment stability
interchange of limits (T2.3)

FIG. 2. *Road map (for multiclass queueing network; T, L: theorem, lemma).*

times, and $\{m_k\}_{k\in\mathcal{K}}$ be a sequence of real numbers with $m_k \geq 1$. Define the regular events as

$$\begin{aligned}\Omega^k(t^*,u^*,m_k) &= \Omega_u^k(t^*,m_k)\cap\Omega_v^k(t^*,m_k)\cap\Omega_X^k(t^*,m_k)\\ &\quad\cap\Omega_E^k(t^*,u^*,m_k)\cap\Omega_S^k(t^*,u^*,m_k)\cap\Omega_\Phi^k(t^*,u^*,m_k),\end{aligned}\tag{3.3}$$

where

$$\Omega_u^k(t^*,m_k) = \bigcap_{r\in\mathcal{R}}\left\{\frac{1}{km_k}u_r^{k,\max}(k^2m_kt^*) \leq \frac{1}{k^{(p^*-2)/2p^*}}\right\},\tag{3.4}$$

$$\Omega_v^k(t^*,m_k) = \bigcap_{r\in\mathcal{R}}\left\{\frac{1}{km_k}v_r^{k,\max}(k^2m_kt^*) \leq \frac{1}{k^{(p^*-2)/2p^*}}\right\},\tag{3.5}$$

$$\begin{aligned}&\Omega_E^k(t^*,u^*,m_k)\\ &\quad=\left\{\sup_{0\leq t\leq kt^*}\sup_{0\leq u\leq u^*}\left|\frac{1}{m_k}(\bar{E}^{o,k}(m_k(t+u))-\bar{E}^{o,k}(m_kt))-\alpha^ku\right|\right.\\ &\qquad\left.\leq\frac{1}{\log k}\right\},\end{aligned}\tag{3.6}$$

$$\begin{aligned}&\Omega_S^k(t^*,u^*,m_k)\\ &\quad=\left\{\sup_{0\leq t\leq kt^*}\sup_{0\leq u\leq u^*}\left|\frac{1}{m_k}(\bar{S}^{o,k}(m_k(t+u))-\bar{S}^{o,k}(m_kt))-\mu u\right|\right.\\ &\qquad\left.\leq\frac{1}{\log k}\right\},\end{aligned}\tag{3.7}$$

$$\Omega_\Phi^k(t^*,u^*,m_k)$$

$$= \left\{ \sup_{0 \le t \le k\kappa_\mu t^*} \sup_{0 \le u \le \kappa_\mu u^*} \left| \frac{1}{m_k} (\bar{\Phi}^k(m_k(t+u)) - \bar{\Phi}^k(m_k t)) - Pu \right| \le \frac{1}{\log k} \right\} \qquad \text{where } \kappa_\mu := 2|\mu|, \tag{3.8}$$

$$\Omega_X^k(t^*, m_k) = \left\{ \sup_{0 \le t \le t^*} \frac{1}{m_k} (|\hat{E}^{o,k}(m_k t)| + |\hat{S}^{o,k}(m_k t)| + |\hat{\Phi}^k(m_k t)|) \le \frac{k}{\log k} \right\}. \tag{3.9}$$

Here, we have introduced a sequence of regular events associated with "nice" sample paths; for example, the fluid-scaled arrival processes lie within a certain range of their means according to the definition of $\Omega_E^k(t^*, u^*, m_k)$. Note that the ranges that bound the sample paths are carefully specified such that the probabilities of these events must approach one at a certain rate as indicated in the following lemma, with the proof deferred to Appendix A.

LEMMA 3.1. *Let t^* and u^* be any positive times. Then the following estimate holds for sufficiently large k (depending on t^* and u^*):*

$$\mathsf{P}(\Omega^k(t^*, u^*, m_k)) \ge 1 - \frac{(\log k)^{p^*+1}}{k^{p^*/2-2}} \qquad \text{for all } m_k \ge 1.$$

Complementarity and oscillation inequality. The required oscillation inequality, given in the lemma below, is a standard result (e.g., [39]). To state the inequality, denote for any RCLL (vector) function $f(u)$ ($u \ge 0$) and any time interval $[s, t]$,

$$\mathsf{Osc}(f(\cdot), [s, t]) = \sup\{|f(u_1) - f(u_2)| : s \le u_1 \le u_2 \le t\}.$$

LEMMA 3.2. *Suppose there exists a constant $\kappa_c > 0$ such that for any RCLL functions, $w(t) = (w_\ell(t))_{\ell \in \mathcal{L}}$, $x(t) = (x_\ell(t))_{\ell \in \mathcal{L}}$ and $y(t) = (y_\ell(t))_{\ell \in \mathcal{L}}$, satisfying*

$$w(t) = w(0) + x(t) + \Psi y(t) \ge 0 \qquad \text{for } t \ge 0;$$

$$y_\ell(t) \text{ is nondecreasing in } t \ge 0, \qquad y_\ell(0) = 0, \qquad \ell \in \mathcal{L};$$

$$y_\ell(t) \text{ cannot increase at time } t \qquad \text{if } w(t) > 0.$$

Then the following oscillation inequalities hold for any $0 \le s \le t$,

$$\mathsf{Osc}(w(\cdot), [s, t]) \quad \text{and} \quad \mathsf{Osc}(y(\cdot), [s, t]) \le \kappa_c(\mathsf{Osc}(x(\cdot), [s, t])). \tag{3.10}$$

(*Recall that the reflection matrix Ψ is assumed to be completely-S.*)

The complementarity property is a part of the dynamics of multiclass queueing networks, that is, the condition in (2.7) or (2.20).

Uniform attraction. The required form of uniform attraction property is slightly more general than the one stated earlier in (2.22)–(2.26). It involves the following more general fluid model that includes the initial residuals:

$$\begin{aligned}\bar{q}(t) &= \bar{q}(0) + \operatorname{diag}(\alpha)\big[et - \bar{u}(1)\big]^+ - \big(I - P^T\big)M^{-1}\big[\bar{d}(t) - \bar{v}(1)\big]^+ \\ &= \bar{q}(0) - \operatorname{diag}(\alpha)\big[et \wedge \bar{u}(1)\big] + \big(I - P^T\big)M^{-1}\big[\bar{d}(t) \wedge \bar{v}(1)\big] \\ &\quad + \alpha t - \big(I - P^T\big)M^{-1}\bar{d}(t),\end{aligned} \tag{3.11}$$

$$\bar{w}(t) = CM\bar{q}(t) \geq 0, \tag{3.12}$$

$$\bar{y}(t) = et - CM\bar{d}(t), \qquad \bar{d}(t) \text{ and } \bar{y}(t) \text{ are nondecreasing,} \tag{3.13}$$

$$\int_0^\infty \bar{w}_\ell(t)\, d\bar{y}_\ell(t) = 0, \qquad \ell \in \mathcal{L}, \tag{3.14}$$

(3.15) additional conditions specifying the HL discipline.

[Recall, in the static buffer priority discipline, the additional condition in (3.15) is the same as the one in (2.27).]

We require (among other things) the following uniform attraction property.

THEOREM 3.3. *Assume the heavy traffic condition in* (2.10) *and the uniform attraction property in* (2.28) *hold.*

(a) *There exist a time $\tau \geq 0$ and a constant $\kappa_w > 0$ that only depend on the network parameters such that for any solution to the fluid model in* (3.11)–(3.15) *satisfying $|\bar{q}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq 1$, the workload $\bar{w}(t)$ is nondecreasing in time $t \geq \tau$ and*

$$\big|\bar{w}(t)\big| \leq \kappa_w, \qquad t \geq 0. \tag{3.16}$$

(b) (*Uniform attraction*) *The uniform attraction property defined through the requirement in* (2.28) *or* (2.29) *holds for the fluid model in* (3.11)–(3.15), *too. Consequently, for any given $\varepsilon > 0$ and some sufficiently large time T* (*depending on ε*), *the following holds for any solution to the fluid model in* (3.11)–(3.15) *satisfying $|\bar{q}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq 1$:*

$$\big|\bar{q}(t) - \Delta_q \bar{w}(t)\big| \leq \varepsilon \qquad \textit{for } t \geq T. \tag{3.17}$$

(c) *If $\bar{q}(0) = \Delta_q w^*$ for some $w^* \geq 0$ and $(\bar{u}(1), \bar{v}(1)) = 0$* (*i.e., the initial state is a fixed-point state*), *then $\bar{q}(t) = \bar{q}(0)$ and $\bar{d}(t) = M\lambda t$ for all $t \geq 0$.*

According to a well-known result (in the proof of Theorem 5.2 of Chen [10]), we know that the fluid model in (3.11)–(3.15) coincides with the standard one in (2.22)–(2.26) after a finite time that depends only on the network parameters. That is, the complications incurred by the initials $\bar{u}(1)$ and $\bar{v}(1)$ in Theorem 3.3 can be mitigated and, therefore, the conclusions in the theorem are immediate.

Uniform continuity. First, assume that the Markovian state descriptor is given as $\Xi^k(t) = (Q^k(t), U^k(t), V^k(t))$, or the diffusion-scaled version $\hat{\Xi}^k(t) = (\hat{Q}^k(t), \hat{U}^k(t), \hat{V}^k(t))$, for ease of exposition. As we have seen in Section 2, this descriptor applies under the static buffer priority, head-of-line proportional processor sharing and queue-ratio disciplines, etc.; refer to [4, 26] for more details on the latter two. For the first-in-first-out discipline, we must append additional information that captures the order of arrivals to $\Xi^k(t)$ to form a complete descriptor; refer to [4], too. Nevertheless, the additional information plays a role in our study only through (various versions of) the corresponding fluid model and the related uniform attraction property, and hence it is sufficient to assume the above descriptor for our purpose.

Denote

$$\begin{aligned} & y^k[= y^k(\omega, \Delta, m_k)] \\ & \quad := \max\biggl(\frac{1}{m_k}|\hat{W}^k(0)| + \sup_{0\le t\le\Delta} \frac{1}{m_k}(|\hat{X}^k(m_k t)| + |\hat{\varepsilon}^k(m_k t)|), \\ & \qquad \frac{1}{m_k}|\hat{\Xi}^k(0)|, 1\biggr), \end{aligned} \tag{3.18}$$

for any time interval $[0, \Delta]$, with $\Delta > 0$, and any sequence of numbers $\{m_k \ge 1; k \in \mathcal{K}\}$. Let $T > 0$ be a fixed time of a certain magnitude (to be specified later). Divide the time interval $[0, \Delta]$ into a total of $\lceil k\Delta/y^k T\rceil$ segments with equal length $y^k m_k T/k$, where $\lceil\cdot\rceil$ denotes the integer ceiling. The jth segment, $j = 0, \dots, \lceil k\Delta/y^k T\rceil - 1$, covers the time interval $[j y^k m_k T/k, (j+1) y^k m_k T/k]$. Note that the last interval (with $j = \lceil k\Delta/y^k T\rceil - 1$) covers a negligible piece of time beyond the right end of $[0, \Delta]$ if $k\Delta/y^k T$ is not an integer. For simplicity, below we shall treat $k\Delta/y^k T$ as an integer so as to omit the ceiling notation. Then, for any $t \in [0, \Delta]$, we can write $t = y^k m_k (jT + u)/k$ for some $j = 0, \dots, k\Delta/y^k T$ and $u \in [0, T]$. Therefore, for $u \in [0, T]$ and $j \le k\Delta/y^k T$, we write

$$\begin{aligned} \frac{1}{y^k m_k}\hat{W}^k(m_k t) &= \frac{1}{y^k m_k}\hat{W}^k\biggl(\frac{j y^k m_k T + y^k m_k u}{k}\biggr) \\ &= \frac{1}{k y^k m_k} W^k(j k y^k m_k T + k y^k m_k u) := \bar{W}^{k,j}(u). \end{aligned} \tag{3.19}$$

The definition for the "hydrodynamic" scaling above, $\bar{W}^{k,j}(u)$, is slightly different from the same notation in our previous papers [40–42] in that new parameters m_k and y^k are introduced into the scaling. The other processes, $\bar{\Xi}^{k,j}(u)$, $\bar{Q}^{k,j}(u)$, $\bar{U}^{k,j}(u)$ and $\bar{V}^{k,j}(u)$, are defined in the same manner.

The uniform continuity property involves approximating the hydrodynamics systems by the fluid model in (3.11)–(3.15).

LEMMA 3.4 (Uniform continuity). *Let M, Δ (and $\bar{\Delta} := \Delta + 1$), and T be any given positive numbers.*

(a) *For any* $\varepsilon > 0$, *there exists* k^* *such that for any* $k \geq k^*$, *the following holds for any* $m_k \geq |\hat{\Xi}^k(0)| \vee 1$, $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, *and* $0 \leq j \leq k\Delta/y^k T$: *if*

$$|\bar{Q}^{k,j}(0)| + |\bar{U}^{k,j}(0)| + |\bar{V}^{k,j}(0)| \leq M, \tag{3.20}$$

then, we can find a fluid model $(\bar{q}(t), \bar{u}(1), \bar{v}(1))$ *satisfying* (3.11)–(3.15) *and* $|\bar{q}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq M$ *such that*

$$\sup_{0 \leq u \leq T} |\bar{Q}^{k,j}(u) - \bar{q}(u)| + |\bar{U}^{k,j}(0) - \bar{u}(1)| + |\bar{V}^{k,j}(0) - \bar{v}(1)| < \varepsilon.$$

(b) *Moreover, the time* T *can be chosen sufficiently long* (*depending on network parameters only*) *such that the following holds for any* $m_k \geq |\hat{\Xi}^k(0)| \vee 1$, $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$ *and* $1 \leq j \leq k\Delta/y^k T$ (*excluding* $j = 0$):

$$\bar{U}^{k,j}(0) \quad \textit{and} \quad \bar{V}^{k,j}(0) \leq \frac{1}{k^{(p^*-1)/2p^*}}.$$

Consequently, for any $\varepsilon > 0$, *there exists* k^* *such that the following holds for any* $m_k \geq |\hat{\Xi}^k(0)| \vee 1$, $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, *and* $k \geq k^*$, $1 \leq j \leq k\Delta/y^k T$, *if*

$$|\bar{Q}^{k,j}(0)| \leq M, \tag{3.21}$$

then we can find a fluid model $(\bar{q}(t), \bar{u}(1) = 0, \bar{v}(1) = 0)$ *satisfying* (3.11)–(3.15) *and* $|\bar{q}(0)| \leq M$ *such that*

$$\sup_{0 \leq u \leq T} |\bar{Q}^{k,j}(u) - \bar{q}(u)| < \varepsilon.$$

Recall, involving $\bar{u}(1)$ and $\bar{v}(1)$ in this set of equations is necessary for handling the more general initial states in the prelimit networks [refer to the property (a) in the above lemma]. Such a uniform continuity property essentially follows the approach of Bramson [4], even though the initial residuals and the additional scaling parameters (m_k and y^k) are *not* considered in his version. The proof is detailed in Appendix B.

3.2. *Bounding workloads over regular events, and pth moment bound of workloads.* Given the above preparations, we follow the hydrodynamic approach to establish the bound for the workload process under the regular events.

LEMMA 3.5. *Consider any time interval* $[0, \Delta]$, *with* $\Delta > 0$. *Let* $\varepsilon > 0$ *be any given* (*small*) *number. Then there exists a sufficiently large* T *such that for sufficiently large* k, *the following results hold for any initial state* $\hat{\Xi}^k(0)$, $m_k \geq |\hat{\Xi}^k(0)| \vee 1$, $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$ (*here* $\bar{\Delta} = \Delta + 1$), *and positive integers* $j = 1, \dots, k\Delta/y^k T$:

(a) (*Uniform attraction*)

$$|\bar{Q}^{k,j}(u) - \Delta_q \bar{W}^{k,j}(u)| \leq \varepsilon \qquad \textit{for all } u \in [0, T]; \tag{3.22}$$

(b) (*Boundedness*)

$$|\bar{W}^{k,j}(u)| \le \kappa \qquad \textit{for all } u \in [0, T], \tag{3.23}$$

where κ is a positive constant that depends only on network parameters (independent of k and ω). In addition, the bound in (3.23) also applies to $j = 0$.

The above lemma is stronger than similar results in the literature for establishing a conventional heavy-traffic theorem (e.g., Lemma 7 of [42]): here the results in (a), (b) hold uniformly on the regular events and allow for additional scaling parameters. [Specifically, (b) is what is needed to bound the pth moment of the workload below, while (a) is an auxiliary result.]

PROOF OF LEMMA 3.5. Let T be any real value satisfying:

$$T \ge T_{\bar{\kappa},\varepsilon/8}, \tag{3.24}$$

where the term on the right-hand side is defined in Theorem 3.3(b). Note that T is large enough so that in the fluid network in Theorem 3.3 (under the heavy traffic condition), the state $\bar{q}(t)$ will be close enough (by an error bound of $\varepsilon/8$) to the fixed-point state, starting from an initial state $(\bar{q}(0), \bar{u}(1), \bar{v}(1))$ that is bounded by $\bar{\kappa}$. Here, ε is given in the current lemma under proof, and $\bar{\kappa}$ is a constant that depends on network parameters only:

$$\kappa = \kappa_w + 3\kappa_c + 1, \qquad \bar{\kappa} = \kappa_\mu \kappa + 1, \tag{3.25}$$

where κ_w and κ_c are given in Theorem 3.3 and Lemma 3.2. The rationale for the choice of κ will become evident shortly.

Step 1. Prove (a), (b) of Lemma 3.5 for $j = 1$.

Let $\varepsilon' > 0$ be any given number. Note that $|\bar{\Xi}^{k,0}(0)| = |\hat{\Xi}^k(0)/y^k| \le 1$ according to the definitions in (3.18), (3.19). By Lemma 3.4, we have for sufficiently large k, and for any initial state $\hat{\Xi}^k(0)$, $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, there exists a fluid model $(\bar{q}(u), \bar{u}(1), \bar{v}(1))$ satisfying (3.11)–(3.15), which may depend on k, $\hat{\Xi}^k(0)$, m_k and ω, such that

$$\begin{aligned} &|\bar{q}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \le 1 \quad \text{and} \\ &\sup_{u\in[0,2T]} \left(|\bar{Q}^{k,0}(u) - \bar{q}(u)| + |\bar{U}^{k,0}(0) - \bar{u}(1)| + |\bar{V}^{k,0}(0) - \bar{v}(1)|\right) < \varepsilon'. \end{aligned} \tag{3.26}$$

Since $T \ge T_{\bar{\kappa},\varepsilon/8}$, applying the uniform attraction property in Theorem 3.3 to the above fluid model $(\bar{q}(u), \bar{w}(u))$ yields:

$$\begin{aligned} &|\bar{q}(u) - \Delta_q \bar{w}(u)| \le \frac{\varepsilon}{8} \qquad \text{for all } u \ge T \quad \text{and} \\ &|\bar{w}(u)| \le \kappa_w\left(|\bar{q}(0)| + |\bar{u}(1)| + |\bar{v}(1)|\right) \le \kappa_w \qquad \text{for all } u \ge 0. \end{aligned} \tag{3.27}$$

Note that $(\bar{Q}^{k,0}(T+u), \bar{W}^{k,0}(T+u)) \equiv (\bar{Q}^{k,1}(u), \bar{W}^{k,1}(u))$ [and $\bar{Q}^{k,0}(0) = \hat{Q}^k(0)/y^k$]; and that

$$\begin{aligned}|\bar{Q}^{k,1}(u) - \Delta_q \bar{W}^{k,1}(u)| \le |\bar{Q}^{k,1}(u) - \bar{q}(T+u)| + |\bar{q}(T+u) - \Delta_q \bar{w}(T+u)| \\ + |\Delta_q| \cdot |\bar{w}(T+u) - \bar{W}^{k,1}(u)|.\end{aligned}$$

Hence, choosing a sufficiently small ε' at the beginning of the proof, the estimate in (3.26), along with (3.27), implies that the conclusion (a) holds with $j = 1$ for sufficiently large k and for all $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$. By (3.26), (3.27) again, we have for all $u \in [0, 2T]$,

$$|\bar{W}^{k,0}(u)| \le |\bar{w}(u)| + \varepsilon' \le \kappa_w + \varepsilon' \le \kappa_w + \varepsilon,$$

and for all $u \in [0, T]$,

$$|\bar{W}^{k,1}(u)| = |\bar{W}^{k,0}(T+u)| \le \kappa_w + \varepsilon \ (\le \kappa). \tag{3.28}$$

That is, the bounding property in (b), for both $j = 0$ and $j = 1$, is satisfied.

Step 2. We now extend the above to $j = 2, \ldots, k\delta/y^k T$. Suppose again, to the contrary, there exists a subsequence $\mathcal{K}_1$ of k such that, for any $k \in \mathcal{K}_1$, at least one of the results in (a), (b) does not hold for some integer $j \in [2, k\delta/y^k T]$ and for some $\hat{\Xi}^k(0)$, $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and sample-path $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$. Let j_k be the smallest positive integer in the interval $[2, k\delta/y^k T]$ such that at least one of the properties in (a), (b) does not hold with the associated $\hat{\Xi}^k(0)$, m_k and ω. To reach a contradiction, in the rest of the proof, we will show that the desired properties in (a), (b) hold for $j = j_k$ for sufficiently large $k \in \mathcal{K}_1$, and indeed for any initial state $\hat{\Xi}^k(0)$, $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$.

Let $\varepsilon' > 0$ be any given number. Following the earlier argument, under the (contradictory) assumption above, the results in (a), (b) hold for $j = 1, \ldots, j_k - 1$, any $\hat{\Xi}^k(0)$, $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, for each $k \in \mathcal{K}_1$. Specifically, for $j = j_k - 1$ (≥ 1), we have

$$|\bar{W}^{k,j_k-1}(0)| \le \kappa \qquad \text{for all } k \in \mathcal{K}_1.$$

This yields the following:

$$\begin{aligned}&|\bar{Q}^{k,j_k-1}(0)| + |\bar{U}^{k,j_k-1}(0)| + |\bar{V}^{k,j_k-1}(0)| \\ &\quad \le \kappa_\mu |\bar{W}^{k,j_k-1}(0)| + 1 \le \bar{\kappa} \qquad \text{for all } k \in \mathcal{K}_1.\end{aligned}$$

By Lemma 3.4(b), we have for any sufficiently large $k \in \mathcal{K}_1$, and for any $\hat{\Xi}^k(0)$, $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, there exists a fluid model $(\bar{q}(u), \bar{w}(u))$ satisfying (3.11)–(3.15) [which may depend on k, $\hat{\Xi}^k(0)$, m_k and ω] such that

$$\sup_{u \in [0,2T]} |\bar{Q}^{k,j_k-1}(u) - \bar{q}(u)| < \varepsilon' \tag{3.29}$$

with $|\bar{q}(0)| \le \bar{\kappa}$. [Here, we know that $|\bar{U}^{k,j_k-1}(0)| + |\bar{V}^{k,j_k-1}(0)| \to 0$ as $k \to 0$, and can set $\bar{u}(1) = \bar{v}(1) = 0$ by Lemma 3.4(b).] Since $T \ge T_{\bar{\kappa},\varepsilon/8}$, applying the uniform attraction property in Theorem 3.3 to the above limit yields:

$$|\bar{q}(u) - \Delta_q \bar{w}(u)| \le \frac{\varepsilon}{8} \qquad \text{for all } u \ge T. \tag{3.30}$$

Note that $(\bar{Q}^{k,j_k-1}(T+u), \bar{W}^{k,j_k-1}(T+u)) \equiv (\bar{Q}^{k,j_k}(u), \bar{W}^{k,j_k}(u))$. Hence, the convergence in (3.29), along with (3.30), implies that (a) holds with $j = j_k$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, for sufficiently large $k \in \mathcal{K}_1$.

Consider any sufficiently large $k \in \mathcal{K}_1$, such that the results in (a) hold for $j = 1, \dots, j_k$ (but it needs not holds for $j = 0$) and for all $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$. Fix any $\hat{\Xi}^k(0)$, $m_k \ge |\hat{\Xi}^k(0)| \vee 1$, and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$. This implies that the processes, $(w(t), x(t), y(t), z(t)) := (\hat{W}^k(m_k t), \hat{X}^k(m_k t), \hat{Y}^k(m_k t), \hat{Z}^k(m_k t))/y^k m_k$, satisfy the specifications in Lemma 3.2 *in the time interval* $t \in [y^k T/k, (j_k y^k T + y^k T)/k]$, which merges all intervals corresponding to $j = 1, \dots, j_k$. Hence, we have for any $t \in [y^k T/k, (j_k y^k T + y^k T)/k]$ $(\subset [0, \Delta])$,

$$\begin{aligned} &\mathsf{Osc}\biggl(\frac{1}{y^k m_k}\hat{W}^k(m_k s), s \in [y^k T/k, t]\biggr) \\ &\quad \le \kappa_c \mathsf{Osc}\biggl(\frac{1}{y^k m_k}[\hat{X}^k(m_k s) + \hat{\varepsilon}^k(m_k s)], s \in [y^k T/k, t]\biggr) = 3\kappa_c. \end{aligned} \tag{3.31}$$

[Here, recall that $\hat{\varepsilon}^k(m_k s) = \hat{Q}^k(m_k s) - \Delta_q \hat{W}^k(m_k s)$, which is a small value in the above estimate according to the conclusion (a).] Consequently, we have the following estimations:

$$\begin{aligned} \biggl|\frac{1}{y^k m_k}\hat{W}^k(m_k t)\biggr| &\le \biggl|\frac{1}{y^k m_k}\hat{W}^k(m_k y^k T/k)\biggr| \\ &\qquad + \mathsf{Osc}\biggl(\frac{1}{y^k m_k}\hat{W}^k(m_k s), s \in [y^k T/k, t]\biggr) \\ &\le \kappa_w + \varepsilon + 3\kappa_c, \end{aligned}$$

where in the second inequality we have also applied the conclusion in (3.28), that is, $|\hat{W}^k(m_k y^k T/k)/y^k m_k| = |\bar{W}^{k,1}(0)| \le \kappa_w + \varepsilon$. Keeping in mind that $\bar{W}^{k,j_k}(u) \equiv \hat{W}^k((j_k y^k m_k T + y^k m_k u)/k)/y^k m_k$, the above implies that (b) holds with $j = j_k$ for sufficiently large $k \in \mathcal{K}_1$. □

Next, we establish the pth moment bound for workloads and queue lengths.

LEMMA 3.6 (Bounded pth moment of workload). *There is a constant $\kappa > 0$ such that for any time $t \ge 0$ and sufficiently large k, the following holds for any initial state $\hat{\Xi}^k(0)$ and any $m_k \ge |\hat{\Xi}^k(0)| \vee 1$:*

$$\mathsf{E} \sup_{0 \le s \le t} \biggl|\frac{1}{m_k}\hat{W}^k(m_k s)\biggr|^p \le \kappa(1 + t^p).$$

PROOF. Following Lemma 3.5, we first bound the workload processes in $\Omega^k(\bar{\Delta}, T, m_k)$. By Lemma 3.5(b), there is a constant κ_1 such that the following holds for sufficiently large k, any initial state $\hat{\Xi}^k(0)$, any $m_k \geq |\hat{\Xi}^k(0)| \vee 1$, and any $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$:

$$\sup_{0\leq t\leq\Delta}\left|\frac{1}{m_k}\hat{W}^k(m_k t)\right|$$
$$\leq \kappa_1 y^k \leq \kappa_1\left(\frac{|\hat{W}^k(0)|}{m_k} + \sup_{0\leq s\leq\Delta}\frac{|\hat{X}^k(m_k s)|}{m_k} + \frac{|\hat{\Xi}^k(0)|}{m_k} + 1\right),$$

where in the second inequality we have used Lemma 3.5(a) to remove the term $\hat{\varepsilon}^k(m_k s)$. Hence, we have

$$\mathsf{E}\left(\sup_{0\leq t\leq\Delta}\frac{1}{m_k}|\hat{W}^k(m_k t)|\right)^p \cdot 1_{\Omega^k(\bar{\Delta},T,m_k)}$$
$$\leq \kappa_1\mathsf{E}(y^k)^p 1_{\Omega^k(\bar{\Delta},T,m_k)} \leq \kappa_1\mathsf{E}(y^k)^p \leq \kappa_2(1+\Delta^p),$$

where the last inequality is proved following the same procedure as in Lemma 9(a) of [43]. Next, we bound the workload in $\Omega \setminus \Omega^k(\bar{\Delta}, T, m_k)$. Pick any positive number α and β satisfying $1/\alpha + 1/\beta = 1$ and in addition $1 < \beta < (p^* - 4)/2p$. Then, for sufficiently large k, we have

$$\mathsf{E}\left(\sup_{0\leq t\leq\Delta}\left|\frac{1}{m_k}\hat{W}^k(m_k t)\right|\right)^p \cdot 1_{\Omega\setminus\Omega^k(\bar{\Delta},T,m_k)}$$
$$\leq \left[\mathsf{E}\left(\sup_{0\leq t\leq\Delta}\left|\frac{1}{m_k}\hat{W}^k(m_k t)\right|\right)^{\alpha p}\right]^{\frac{1}{\alpha}} \cdot [\mathsf{E}(1_{\Omega\setminus\Omega^k(\bar{\Delta},T,m_k)})^\beta]^{\frac{1}{\beta}}$$
$$\leq \left[\mathsf{E}\kappa_2\left(1 + \left|\frac{1}{km_k}E^{o,k}(k^2 m_k\Delta)\right|^{\alpha p}\right)\right]^{\frac{1}{\alpha}} \cdot [\mathsf{P}(\Omega \setminus \Omega^k(\bar{\Delta}, T, m_k))]^{\frac{1}{\beta}}$$
$$\leq [\kappa_3(1 + (k\Delta)^{\alpha p})]^{\frac{1}{\alpha}} \cdot \left[\frac{(\log k)^{p^*+1}}{k^{p^*/2-2}}\right]^{\frac{1}{\beta}}$$
$$\leq \kappa_4(1+\Delta^p).$$

Here, we have applied Hölder's inequality and Lemma 3.1 in the first and forth inequalities, respectively, and have taken into account $p^* > 2(p+2)$ and our choice of β in the last inequality. To see the second inequality, we note the following estimate (for some constants κ_2' and κ_2''):

$$\frac{1}{m_k}|\hat{W}_r^k(m_k t)| \leq \frac{\kappa_2'}{m_k}|\hat{Q}_r^k(m_k t)| \leq \frac{\kappa_2'}{m_k}|\hat{Q}_r^k(0)| + \sum_r \frac{\kappa_2'}{km_k}E_r^k(k^2 m_k t)$$
$$\leq \kappa_2' + \sum_r \frac{\kappa_2'}{km_k}[1 + E_r^{o,k}(k^2 m_k t)]$$

$$\le \kappa_2' + \frac{\kappa_2'}{km_k} R + \sum_r \frac{\kappa_2'}{km_k} E_r^{o,k}(k^2 m_k t)$$

$$\le \kappa_2'' \biggl(1 + \frac{1}{km_k} |E^{o,k}(k^2 m_k t)| \biggr).$$

Finally, the above two estimates lead to the lemma, since the time Δ is arbitrarily given in Lemma 3.5. □

Bounding the pth moment of the workload in the above lemma is a crucial step in establishing (in the next subsection) the uniform integrability of the workload processes, and subsequently its uniform pth moment stability and tightness. To do so, we cannot simply follow similar cases in the literature, for example, Dai ([15], Lemma 4.5) and Dai and Meyn ([19], Lemma 5.2), where the fluid-scaled arrival processes serve as the bound for the workload, and thus leading to the required uniform integrability property directly. For example, in Dai and Meyn [19], to justify the existence of the pth stationary moment for a single multiclass queue network (say, the kth network), a moment bound for the queue-length process can be derived by discarding the service and routing processes in (2.3): for any sufficiently large initial state $x = \Xi^k(0)$ $(= (Q^k(0), U^k(0), V^k(0)))$,

$$\mathsf{E} \sup_{0 \le s \le t} \left| \frac{1}{|x|} Q^k(|x|s) \right|^p \le \mathsf{E} \sup_{0 \le s \le t} \left| \frac{1}{|x|} \bigl(Q^k(0) + E^k(|x|s) \bigr) \right|^p \le \kappa (1 + t^p).$$

However, under diffusion scaling (which is needed in our setting), the arrival processes become unbounded as $k \to \infty$ and cannot serve the same purpose to obtain the bound in Lemma 3.6. To overcome this difficulty turns out to be a major effort, and our approach is to identify the regular events and characterize the hydrodynamics of the networks under these events as summarized in Lemmas 3.1 and 3.5, which has provided proper preparations for the proof of Lemma 3.6.

3.3. *Uniform stability, uniform integrability, uniform pth moment stability, tightness and interchange of limits.* As demonstrated in [43], the uniform stability property is a general property that applies to a wide class of stochastic processing networks, including the multiclass queueing networks under study.

We state the property for reference here. For each k, denote $(\bar{q}^k(t), \bar{w}^k(t))$ as the fluid model corresponding to the kth (original, discrete) network $(Q^k(t), W^k(t))$. It satisfies the relationship in (3.11)–(3.15) with the index k properly appended, and roughly speaking is derived as the limit of $(Q^k(nt), W^k(nt))/n$ as $n \to \infty$. Denote the diffusion-scaled version as $(\hat{q}^k(t), \hat{w}^k(t)) := (\bar{q}^k(k^2 t), \bar{w}^k(k^2 t))/k$.

LEMMA 3.7 (Uniform stability). *Assume that the DCP $\hat{w}(t)$ defined through (2.34)–(2.36) is stable. Then there exists a time $t_0 > 0$ such that for any sufficiently large k, the (diffusion-scaled version of) fluid models $(\hat{q}^k(t), \hat{w}^k(t))$, with $|\hat{w}^k(0)| \le 1$ [or equivalently $|\hat{q}^k(0)| \le 1$], satisfies the following*:

$$\hat{w}^k(t) = 0, \qquad t \ge t_0. \tag{3.32}$$

Now, the rest of the building blocks in Figure 2, including the uniform integrability, uniform pth moment stability, tightness and finally the interchange of limits, follow from the approach in [43]. This is because the technical developments for these blocks are independent of the specific network structure (either the resource-sharing network in [43] or the multiclass queueing network here), given the pth moment bound of workloads and the uniform stability. We outline the main steps as follows.

LEMMA 3.8 (Uniform integrability). (a) *Let* $\{m_i; i = 1, 2, \dots\}$ *be a sequence of number such that* $m_i \to \infty$ *as* $i \to \infty$; *and let* $\{x^i \in \mathcal{X}; i = 1, 2, \dots\}$ *be a sequence of initial states such that* $|x^i| \le m_i$ *for all* i. *Then, for any given* $t \ge 0$, *and a fixed, sufficiently large* k, $\{|\hat{W}^k(m_i t; x^i)/m_i|^p\}$ *is uniformly integrable* (*w.r.t.* i).

(b) *Let* $\{m_k\}$ *be a sequence of numbers such* $m_k \to \infty$ *as* $k \to \infty$, *and assume that the sequence of initial states satisfies* $|\hat{\Xi}^k(0)| \le m_k$. *Then, for any time* $t \ge 0$, $\{|\hat{W}^k(m_k t)/m_k|^p\}_k$ *is uniformly integrable* (*w.r.t.* k).

PROOF. The proofs of (a) and (b) are similar, and we prove (b) only. Pick any constant p' satisfying $p < p' < p^*/2 - 2$. Clearly, Lemma 3.6 holds with p replaced by p', and hence we have, for some constant κ' and sufficiently large k,

$$\mathsf{E} \sup_{0 \le s \le t} \left| \frac{1}{m_k} \hat{W}^k(m_k s) \right|^{p'} \le \kappa'(1 + t^{p'}),$$

which implies that $\{|\hat{W}^k(m_k t)/m_k|^p\}_k$ is uniformly integrable. □

THEOREM 3.9. *Assume the stability of the deterministic DCP in* (2.34)–(2.36) *and the* p^**th moment condition* [*refer to* (2.38), *adapted to MCQ*].

(a) (*Uniform pth moment stability*) *There exists* t_0 *such that the following holds for all* $t \ge t_0$:

$$\lim_{|x| \to \infty} \sup_k \mathsf{E} \frac{1}{|x|^p} |\hat{W}^k(|x|t; x)|^p = 0, \tag{3.33}$$

$$\lim_{|x| \to \infty} \sup_k \mathsf{E} \frac{1}{|x|^p} |\hat{Q}^k(|x|t; x)|^p = 0. \tag{3.34}$$

(b) (*Tightness*) *The sequence of stationary distributions*, $\{\hat{\pi}^k\}$, *is tight on* $\mathcal{X}$. *Furthermore, if* $p \ge 2$, *then* $\sup_k \mathsf{E}_{\hat{\pi}^k} |\hat{\Xi}^k(0)|^{p-1} < \infty$.

Part (a) of the above theorem follows the same proof as that of Proposition 11 in [43], given the uniform integrability properties in Lemma 3.8 and the uniform stability property in Lemma 3.7; part (b) can be proved making use of the results in part (a) and Lemma 3.6, and following the approaches developed in Budhiraja and Lee [9] and Dai and Meyn [19], as well as in [43]. The detailed proof is thus omitted.

Finally, given the tightness property in Theorem 3.9, the proof of our main result here, Theorem 2.3, is identical to the proof of Theorem 4 in [43], which is a modification of the standard argument leading to the interchange of limits from the tightness in [9, 25, 26, 29, 30].

4. Concluding remarks. As mentioned above and recognized by other authors as well, the key step in justifying the interchange of limits is to bound the pth moment of the workload process. In [26], this step is accomplished through another condition, that the fluid model converges at a linear rate. In our own recent work, [43], this step is carried out through a bounded workload condition. Verifying either condition requires effort, as evident from the specific examples studied in both [26] and [43]. Here, we have demonstrated that, for multiclass queueing networks, the interchange of limits is justified and the only required condition is the moment conditions on primitives, which either holds automatically or is easy to verify.

This type of interchange of limits, under moment conditions on primitives only, can be extended to other stochastic processing networks, in particular the resource-sharing networks as demonstrated in our previous study [43] (albeit requiring a moment condition on *workload*). Specifically, the results concerning edges II and III in Figure 1 can be established, quite routinely, under the stability of a (single) deterministic DCP (that corresponds to the diffusion limit). The key (that requires effort but can be done) is to establish the complementarity property, that is, the condition in (2.7) or (2.20), which automatically holds in multiclass queueing networks (in fact, is part of the dynamics).

APPENDIX A: PROOF OF LEMMA 3.1

The key to proving the lemma is to combine the probability bounds on the various events, which we first construct below. To do so, we need certain estimates on the moments of renewal processes, which are collected in Section A.1 after the proof.

For ease of reading, we prove the lemma for the case $m_k = 1$ for all k. For the general case, simply replace the scaling parameters (not the indices) k and k^2, at relevant places, with km_k and $k \cdot km_k$, respectively, and use the condition $m_k \geq 1$ to remove the parameter m_k in the conclusion. We also omit the arguments t^*, u^* and m_k associated with the events to lighten up notation.

Probability bounds for Ω_u^k and Ω_v^k. We estimate the probability bound for Ω_u^k following the approach in the proof of Lemma 5.1 in Bramson [4]. The bound for Ω_v^k is similar, and hence is omitted.

Denote for each r and k,

$$U_r^k(i) = \sum_{i'=2}^{i} u_r^k(i'), \qquad i \geq 2.$$

Observe from the definition of Ω_u^k that it is sufficient to estimate the probability bound for the event $\{u_r^{k,\max}(k^2t^*)/k \leq 1/k^{(p^*-2)/2p^*}\}$ as the number of job classes is finite. Consider any fixed r, and pick a sufficiently large constant $\kappa_1 > 0$. We have

$$\begin{aligned}
&\mathsf{P}\left\{\frac{1}{k}u_r^{k,\max}(k^2t^*) > \frac{1}{k^{(p^*-2)/2p^*}}\right\} \\
&\quad \leq \mathsf{P}\{U_r^k(\lfloor\kappa_1 k^2 t^*\rfloor) \leq k^2t^*\} \\
&\qquad + \mathsf{P}\left(\left\{\frac{1}{k}u_r^{k,\max}(k^2t^*) > \frac{1}{k^{(p^*-2)/2p^*}}\right\} \cap \{U_r^k(\lfloor\kappa_1 k^2 t^*\rfloor) > k^2t^*\}\right) \\
&\quad := F_1 + F_2.
\end{aligned}$$

The first term F_1 is estimated by applying Lemma A.2 and the Markov inequality as follows:

$$\begin{aligned}
F_1 &\leq \mathsf{P}\left\{\frac{U_r^k(\lfloor\kappa_1 k^2 t^*\rfloor)}{\kappa_1 k^2 t^*} - \frac{1}{\lambda_r^k} \leq \frac{1}{\kappa_1} - \frac{1}{\lambda_r^k}\right\} \\
&\leq \mathsf{P}\left\{\left|U_r^k(\lfloor\kappa_1 k^2 t^*\rfloor) - \frac{1}{\lambda_r^k}(\kappa_1 k^2 t^*)\right| \geq \left(\frac{1}{\lambda_r^k} - \frac{1}{\kappa_1}\right)\kappa_1 k^2 t^*\right\} \leq \frac{\kappa_2}{k^{p^*}}.
\end{aligned}$$

In the event in the term F_2, $u_r^{k,\max}(k^2t^*)$ will be selected among $\{u_r^k(i) : i = 2, \ldots, \lfloor\kappa_1 k^2 t^*\rfloor\}$. Hence, we have

$$\begin{aligned}
F_2 &\leq \mathsf{P}\left(\bigcup_{i=2}^{\lfloor\kappa_1 k^2 t^*\rfloor}\left[\left\{\frac{1}{k}u_r^k(i) > \frac{1}{k^{(p^*-2)/2p^*}}\right\} \cap \{U_r^k(\lfloor\kappa_1 k^2 t^*\rfloor) > k^2t^*\}\right]\right) \\
&\leq \sum_{i=2}^{\lfloor\kappa_1 k^2 t^*\rfloor} \mathsf{P}\left\{\frac{1}{k}u_r^k(i) > \frac{1}{k^{(p^*-2)/2p^*}}\right\} \\
&= (\lfloor\kappa_1 k^2 t^*\rfloor - 1)\mathsf{P}\left\{\frac{1}{k}u_r^k(2) > \frac{1}{k^{(p^*-2)/2p^*}}\right\} \\
&\leq \lfloor\kappa_1 k^2 t^*\rfloor \mathsf{E}[u_r^k(2)]^{p^*}\left(\frac{k^{(p^*-2)/2p^*}}{k}\right)^{p^*} \leq \frac{\kappa_1 t^* \mathsf{E}[u_r^k(2)]^{p^*}}{k^{p^*/2-1}}.
\end{aligned}$$

The above estimates yield that there is a constant κ_3, independent of k, such that for sufficiently large k,

$$\mathsf{P}\left\{\frac{1}{k}u_r^{k,\max}(k^2t^*) > \frac{1}{k^{(p^*-2)/2p^*}}\right\} \leq \frac{\kappa_3}{k^{p^*/2-1}}.$$

Summing up the above over all r, we have for sufficiently large k,

$$\mathsf{P}(\Omega \setminus \Omega_u^k) \le \frac{\kappa_4}{k^{p^*/2-1}}.$$

Probability bound for Ω_X^k. Observe that the additional process $\Phi_{rs}^k(n)$ is also a renewal process whose interarrival times follow a geometric distribution and, therefore, have all moments. Hence, this process satisfies the p^*-moment condition trivially (and indeed we have implicitly ignored it when we introduce the p^*-moment condition for the multiclass queueing network). Indeed, for any $p' > 2$, the following holds for some constant $\kappa > 0$ and for all $t \ge 0$,

$$\mathsf{E} \sup_{0\le s\le t} \sum_{r,r'\in\mathcal{R}} |\hat{\Phi}_{r,r'}^k(s)|^{p'} \le \kappa\bigl(1 + t^{p'/2}\bigr). \tag{A.1}$$

Note that there exists a constant κ_1 such that

$$\begin{aligned}&\Bigl(\sup_{0\le s\le t} \bigl(|\hat{E}^{o,k}(s)| + |\hat{S}^{o,k}(s)| + |\hat{\Phi}^k(s)|\bigr)\Bigr)^{p^*}\\ &\qquad \le \kappa_1 \sup_{0\le s\le t} \sum_{r\in\mathcal{R}} \bigl(|\hat{E}_r^{o,k}(s)|^{p^*} + |\hat{S}_r^{o,k}(s)|^{p^*} + |\hat{\Phi}^k(s)|^{p^*}\bigr).\end{aligned}$$

Applying the above inequality, the p^*th moment condition in (2.40) [plus (A.1)], Markov inequality and Lemma A.3, we have the following estimation:

$$\begin{aligned}\mathsf{P}(\Omega \setminus \Omega_X^k) &\le \mathsf{E}\Bigl(\sup_{0\le t\le t^*} \bigl(|\hat{E}^{o,k}(t)| + |\hat{S}^{o,k}(t)| + |\hat{\Phi}^k(s)|\bigr)\Bigr)^{p^*} \frac{(\log k)^{p^*}}{k^{p^*}}\\ &\le \kappa_2\bigl(1 + (t^*)^{p^*/2}\bigr)\frac{(\log k)^{p^*}}{k^{p^*}}.\end{aligned}$$

Probability bounds for Ω_E^k, Ω_S^k *and* Ω_Φ^k. We estimate the probability bound for Ω_E^k only, since the bounds for Ω_S^k and Ω_Φ^k are similar.

First, we show that for any positive constant α and for some positive constant κ_1, the following holds for any $r \in \mathcal{R}$, $t \in [0, kt^*]$ and $u \in [0, u_1^*]$ (where $u_1^* := u^* + t^*$), and for sufficiently large k (depending only on network parameters, u^* and t^*):

$$\begin{aligned}J &:= \mathsf{P}\biggl(\biggl\{\bigl|\bar{E}_r^{o,k}(t+u) - \bar{E}_r^{o,k}(t) - \lambda_r^k u\bigr| > \frac{1}{2\alpha\log k}\biggr\} \cap \Omega_u^k\biggr)\\ &\le \kappa_1 \frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}.\end{aligned} \tag{A.2}$$

Note that in the above probability, there is no supremum operator on the event set and the time variable u may take any value over a longer interval $[0, u_1^*]$, in contract to the event Ω_E^k defined in (3.6). Write the term involving the arrival process in (A.2) as

$$\begin{aligned}
&|\bar{E}_r^{o,k}(t+u) - \bar{E}_r^{o,k}(t) - \lambda_r^k u| \\
&\qquad \le |\bar{E}_r^{o,k}(t+\bar{U}_r^k(t)+u) - \bar{E}_r^{o,k}(t+\bar{U}_r^k(t)) - \lambda_r^k u| \\
&\qquad\quad + |\bar{E}_r^{o,k}(t+\bar{U}_r^k(t)+u) - \bar{E}_r^{o,k}(t+u)| + |\bar{E}_r^{o,k}(t+\bar{U}_r^k(t)) - \bar{E}_r^{o,k}(t)|.
\end{aligned}$$

Observe in the first term at the right-hand side of the above that $J_1 := (\bar{E}_r^{o,k}(t+\bar{U}_r^k(t)+u) - \bar{E}_r^{o,k}(t+\bar{U}_r^k(t)) - \lambda_r^k u)$ and $(\bar{E}_r^{o,k}(u) - \lambda_r^k u)$ are equal in distribution. This is because that the time $t+\bar{U}_r^k(t)$ is the arrival time of a class-r job and the process $\bar{E}_r^{o,k}$ is therefore renewed at that time. By the definition of $\bar{U}_r^k(t)$, the third term is equal to $1/k$. For the middle term, we restrict our attention to $\omega \in \Omega_u^k$, which implies $\bar{U}_r^k(t) \le 1/k^{(p^*-2)/2p^*}$. Then we have the following estimate on this term:

$$\begin{aligned}
0 &\le \bar{E}_r^{o,k}(t+u+\bar{U}_r^k(t)) - \bar{E}_r^{o,k}(t+u) \\
&= [\bar{E}_r^{o,k}(t+u+\bar{U}_r^k(t+u) + (\bar{U}_r^k(t) - \bar{U}_r^k(t+u))) \\
&\quad - \bar{E}_r^{o,k}(t+u+\bar{U}_r^k(t+u))] \\
&\quad + [\bar{E}_r^{o,k}(t+u+\bar{U}_r^k(t+u)) - \bar{E}_r^{o,k}(t+u)] \\
&\le \left[\bar{E}_r^{o,k}\left(t+u+\bar{U}_r^k(t+u) + \frac{1}{k^{(p^*-2)/2p^*}}\right) - \bar{E}_r^{o,k}(t+u+\bar{U}_r^k(t+u))\right] + \frac{1}{k} \\
&:= J_2 + \frac{1}{k}.
\end{aligned}$$

Similar to the term J_1 above, the term inside the square bracket (denoted J_2) is equal to $\bar{E}_r^{o,k}(1/k^{(p^*-2)/2p^*})$ in distribution. Putting the above together yields the estimate,

$$|\bar{E}_r^{o,k}(t+u) - \bar{E}_r^{o,k}(t) - \lambda_r^k u| \le J_1 + J_2 + \frac{2}{k}, \qquad t \in [0, kt^*], u \in [0, u_1^*],$$

and consequently,

$$\begin{aligned}
J &\le \mathsf{P}\left(\left\{|J_1| + |J_2| + \frac{2}{k} > \frac{1}{2\alpha \log k}\right\} \cap \Omega_u^k\right) \\
&\le \mathsf{P}\left\{|\bar{E}_r^{o,k}(u) - \lambda_r^k u| > \frac{1}{4\alpha \log k} - \frac{1}{k}\right\} \\
&\quad + \mathsf{P}\left\{\bar{E}_r^{o,k}(1/k^{(p^*-2)/2p^*}) > \frac{1}{4\alpha \log k} - \frac{1}{k}\right\} \\
&\le \frac{\mathsf{E}|\bar{E}_r^{o,k}(u) - \lambda_r^k u|^{p^*}}{(1/4\alpha \log k - 1/k)^{p^*}} + \frac{\mathsf{E}\bar{E}_r^{o,k}(1/k^{(p^*-2)/2p^*})^{p^*}}{(1/4\alpha \log k - 1/k)^{p^*}}
\end{aligned}$$

$$\leq \frac{\kappa'(1/k^{p^*})(1+(ku_1^*)^{p^*/2})}{(1/4\alpha\log k - 1/k)^{p^*}} + \frac{\kappa''(1/k^{p^*})(k/k^{(p^*-2)/2p^*})^{p^*}}{(1/4\alpha\log k - 1/k)^{p^*}}$$

$$\leq \kappa_1 \frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}},$$

where the second last inequality is due to Lemmas A.3 and A.1.

Denote for any r, k and $t \geq 0$,

$$\xi_r^k(t) = \bar{E}_r^{o,k}(t) - \lambda_r^k t, \qquad \tau_r^k(t) = \inf\{u \geq 0 : |\xi_r^k(t+u) - \xi_r^k(t)| > 1/\alpha\log k\}.$$

Clearly, we can write the event:

$$\left\{\sup_{0\leq u\leq u_1^*} |\bar{E}_r^{o,k}(t+u) - \bar{E}_r^{o,k}(t) - \lambda_r^k u| > \frac{1}{\alpha\log k}\right\}$$

$$= \left\{\sup_{0\leq u\leq u_1^*} |\xi_r^k(t+u) - \xi_r^k(t)| > \frac{1}{\alpha\log k}\right\} = \{\tau_r^k(t) \leq u_1^*\}.$$

Following the argument in Bramson ([4], Proposition 4.2), we evaluate the following probability:

$$\mathsf{P}\left\{\tau_r^k(t) \leq u_1^*, |\xi_r^k(t+u_1^*) - \xi_r^k(t+\tau_r^k(t))| \leq \frac{1}{2\alpha\log k}, \omega \in \Omega_u^k\right\}$$

$$= \mathsf{P}\{\tau_r^k(t) \leq u_1^*, \omega \in \Omega_u^k\}$$

$$\times \mathsf{P}\left(\left\{|\xi_r^k(t+u_1^*) - \xi_r^k(t+\tau_r^k(t))| \leq \frac{1}{2\alpha\log k}\right\} \Big| \{\tau_r^k(t) \leq u_1^*, \omega \in \Omega_u^k\}\right)$$

$$\geq \mathsf{P}\{\tau_r^k(t) \leq u_1^*, \omega \in \Omega_u^k\}\left(1 - \kappa_1 \frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}\right),$$

where the inequality is due to (A.2) along with routine conditioning arguments. On the other hand, using (A.2) again,

$$\mathsf{P}\left\{\tau_r^k(t) \leq u_1^*, |\xi_r^k(t+u_1^*) - \xi_r^k(t+\tau_r^k(t))| \leq \frac{1}{2\alpha\log k}, \omega \in \Omega_u^k\right\}$$

$$\leq \mathsf{P}\left\{|\xi_r^k(t+u_1^*) - \xi_r^k(t)| > \frac{1}{2\alpha\log k}, \omega \in \Omega_u^k\right\} \leq \kappa_1 \frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}.$$

The above two imply

$$\mathsf{P}\{\tau_r^k(t) \leq u_1^*, \omega \in \Omega_u^k\} \leq \kappa_1 \frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}\left(1 - \kappa_1 \frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}\right)^{-1} \leq \kappa_2 \frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}},$$

for sufficiently large k. Using the above result, we bound the following probability:

$$\mathsf{P}\left\{\sup_{0\leq t\leq kt^*} \sup_{0\leq u\leq u^*} |\bar{E}_r^{o,k}(t+u) - \bar{E}_r^{o,k}(t) - \lambda_r^k u| > \frac{2}{\alpha\log k}, \omega \in \Omega_u^k\right\}$$

$$
\begin{aligned}
&= \mathsf{P}\Biggl(\bigcup_{j=0}^{k-1}\Biggl\{\sup_{jt^*\le t\le (j+1)t^*}\ \sup_{0\le u\le u^*}\bigl|\bar{E}_r^{o,k}(t+u)-\bar{E}_r^{o,k}(t)-\lambda_r^k u\bigr| \\
&\qquad > \frac{2}{\alpha\log k}, \omega\in\Omega_u^k\Biggr\}\Biggr) \\
&\le \sum_{j=0}^{k-1}\mathsf{P}\Biggl\{\sup_{jt^*\le t\le (j+1)t^*}\ \sup_{0\le u\le u^*}\bigl(\bigl|\bar{E}_r^{o,k}(t+u)-\bar{E}_r^{o,k}(jt^*)-\lambda_r^k(t+u-jt^*)\bigr| \\
&\qquad + \bigl|\bar{E}_r^{o,k}(t)-\bar{E}_r^{o,k}(jt^*)-\lambda_r^k(t-jt^*)\bigr|\bigr) > \frac{2}{\alpha\log k}, \omega\in\Omega_u^k\Biggr\} \\
&\le \sum_{j=0}^{k-1}\mathsf{P}\Biggl\{2\sup_{0\le u\le u_1^*}\bigl|\bar{E}_r^{o,k}(jt^*+u)-\bar{E}_r^{o,k}(jt^*)-\lambda_r^k u\bigr| > \frac{2}{\alpha\log k}, \omega\in\Omega_u^k\Biggr\} \\
&= \sum_{j=0}^{k-1}\mathsf{P}\bigl\{\tau_r^k(jt^*)\le u_1^*, \omega\in\Omega_u^k\bigr\} \le k\kappa_2\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}} \le \kappa_2\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-2}}.
\end{aligned}
$$

With a carefully chosen constant α, the above inequality implies the following immediately:

$$\mathsf{P}\bigl((\Omega\setminus\Omega_E^k)\cap\Omega_u^k\bigr)\le\kappa_3\frac{(\log k)^{p^*}}{k^{p^*/2-2}}.$$

Combined with the probability bound for the event Ω_u^k, the above further implies

$$\mathsf{P}\bigl(\Omega\setminus\Omega_E^k\bigr)\le\kappa_4\frac{(\log k)^{p^*}}{k^{p^*/2-2}}.$$

A.1. Some results on moments of renewal processes. This section collects some independent results on the moments of renewal processes, which are used in the above estimates.

Let X_i, $i=1,2,\dots$, be identically and independently distributed nonnegative random variables, with $\mathsf{E}X_1=\mu>0$. Let $S_n=\sum_{i=1}^n X_i$ $(S_0=0)$, and $Y_t=\max\{n : S_n\le t\}$.

LEMMA A.1. *For any $r>0$, there exists a constant $a_r>0$ (depending on r and the distribution of X_1 only) such that*

$$\mathsf{E}Y_t^r\le a_r\bigl(1+t^r\bigr),\qquad t\ge 0. \tag{A.3}$$

The lemma is a direct result of the strong law of counting (renewal) process (e.g., Theorem 5.1 in Chapter 2 of Gut [27]), and hence its proof is omitted. Note that the lemma requires the existence of the first moment of X_i only.

LEMMA A.2. *Assume* $\mathsf{E}X_i^r < \infty$ *for some* $r \geq 2$. *Then there exists a constant* b_r (*depending on* r *only*) *such that*

$$\mathsf{E}|S_n - n\mu|^r \leq b_r \mathsf{E}|X_1 - \mu|^r n^{\frac{r}{2}},$$

$$\mathsf{E}\Big(\max_{1\leq i\leq n} |S_i - i\mu|\Big)^r \leq \left(\frac{r}{r-1}\right)^r b_r \mathsf{E}|X_1 - \mu|^r n^{\frac{r}{2}}.$$

The first inequality can be found from Gut ([27], page 169), and the second follows from L^p maximum inequality (e.g., Theorem 5.4.3 of Durrett [22]).

LEMMA A.3. *Let* $r > p \geq 2$, *and assume* $\mathsf{E}X_i^r < \infty$. *Then there exists a constant* c *such that for all* $t \geq 0$,

$$\mathsf{E}\Big(\sup_{0\leq s\leq t} |Y_s - \mu^{-1}s|\Big)^p \leq c\big(1 + t^{\frac{p}{2}}\big).$$

This lemma, to the best of our knowledge, first appeared as Theorem 4 of Krichagina and Taksar [31], with a long proof. Here we show, it follows rather quickly from Lemmas A.1 and A.2.

PROOF OF LEMMA A.3. Note that

$$Y_t - \mu^{-1}t = -\mu^{-1}\big(S_{Y_t+1} - \mu(Y_t + 1)\big) - 1 + \mu^{-1}(S_{Y_t+1} - t),$$

and

$$\begin{aligned} S_{Y_t+1} - t \leq S_{Y_t+1} - S_{Y_t} &= \big(S_{Y_t+1} - \mu(Y_t+1)\big) - (S_{Y_t} - \mu Y_t) + \mu \\ &\leq 2 \sup_{0\leq s\leq t} \big|S_{Y_s+1} - \mu(Y_s+1)\big| + \mu. \end{aligned}$$

Hence, we have

$$\sup_{0\leq s\leq t} \big|Y_s - \mu^{-1}s\big| \leq 3\mu^{-1} \sup_{0\leq s\leq t} \big|S_{Y_s+1} - \mu(Y_s+1)\big|. \tag{A.4}$$

Applying Lemma A.2, we have

$$\begin{aligned} &\mathsf{E}\Big(\sup_{0\leq s\leq t} \big|S_{Y_s+1} - \mu(Y_s+1)\big|^p \cdot 1_{\{Y_t < 2\mu^{-1}t+2\}}\Big) \\ &\qquad \leq \mathsf{E}\Big(\sup_{0\leq i\leq 2\mu^{-1}t+3} |S_i - \mu i|^p\Big) \leq c_1'\big(1 + t^{\frac{p}{2}}\big). \end{aligned} \tag{A.5}$$

From (A.4) and (A.5), we have

$$\mathsf{E}\Big(\sup_{0\leq s\leq t} |Y_s - \mu^{-1}s|^p \cdot 1_{\{Y_t < 2\mu^{-1}t+2\}}\Big) \leq c_1\big(1 + t^{\frac{p}{2}}\big). \tag{A.6}$$

On the other hand, denote $\alpha = r/(r-p)$ and $\beta = r/p$, which ensures $1/\alpha + 1/\beta = 1$. Then we have

$$
\begin{aligned}
&\mathsf{E}\Big(\sup_{0\le s\le t} |Y_s - \mu^{-1}s|^p \cdot 1_{\{Y_t \ge 2\mu^{-1}t+2\}}\Big)\\
&\qquad \le \mathsf{E}\big((Y_t^p + (\mu^{-1}t)^p)\cdot 1_{\{Y_t\ge 2\mu^{-1}t+2\}}\big)\\
&\qquad \le \big(\mathsf{E}(Y_t^p + (\mu^{-1}t)^p)^\alpha\big)^{\frac{1}{\alpha}}\big(\mathsf{E}(1_{\{Y_t\ge 2\mu^{-1}t+2\}})^\beta\big)^{\frac{1}{\beta}}\\
&\qquad \le \big(c_2'(1+t^{\alpha p})\big)^{\frac{1}{\alpha}}\big(\mathsf{P}\{Y_t \ge 2\mu^{-1}t+2\}\big)^{\frac{1}{\beta}} \qquad \text{(A.7)}\\
&\qquad \le c_2''(1+t^p)\big(\mathsf{P}\{S_{\lfloor 2\mu^{-1}t+2\rfloor} \le t\}\big)^{\frac{1}{\beta}}\\
&\qquad \le c_2''(1+t^p)\big(\mathsf{P}\{|S_{\lfloor 2\mu^{-1}t+2\rfloor} - \mu\lfloor 2\mu^{-1}t+2\rfloor| \ge \mu + t\}\big)^{\frac{1}{\beta}}\\
&\qquad \le c_2''(1+t^p)\bigg(\frac{b_r \mathsf{E}|X_1-\mu|^r \lfloor 2\mu^{-1}t+2\rfloor^{r/2}}{(\mu+t)^r}\bigg)^{\frac{1}{\beta}} \le c_2\big(1+t^{\frac{p}{2}}\big).
\end{aligned}
$$

Here, we have applied Lemmas A.1 and A.2 in the third and the sixth inequalities, respectively. Finally, the desired result is implied by the inequalities in (A.6) and (A.7). □

We remark that rigorously speaking, the inequalities in (2.39), (2.40) hold for any $p' < p^*$ (instead of p^*) according to the above lemma. Nevertheless, this will not affect any result in the paper, if we choose $p' > 2(p+2)$; and to avoid introducing the annoying extra parameter (p'), we write p^* in these inequalities directly.

APPENDIX B: PROOF OF LEMMA 3.4

To prove the lemma, we first need some preliminary results.

As an abstraction of the fluid-scaled prelimit networks $(\bar{Q}^{k,j}(u), \bar{W}^{k,j}(u))$, we consider the following set of equations on $(q(t), w(t), x(t), u(1), v(1))$:

$$
\begin{aligned}
q(t) = {}& q(0) - \operatorname{diag}(\alpha)[et \wedge u(1)] + (I - P^T)M^{-1}[d(t)\wedge v(1)]\\
&+ x(t) + \alpha t - (I-P^T)M^{-1}d(t),
\end{aligned} \tag{B.1}
$$

$$\bar{w}(t) = CMq(t) \ge 0, \tag{B.2}$$

$$\bar{y}(t) = et - CMd(t), \qquad d(t) \text{ and } y(t) \text{ are nondecreasing}, \tag{B.3}$$

$$\int_0^\infty w_\ell(t)\,dy_\ell(t) = 0, \qquad \ell \in \mathcal{L}, \tag{B.4}$$

(B.5) additional conditions specifying the HL discipline.

In the above, $q(t)$ and $w(t)$ are an R-dimensional nonnegative vector functions of time $t \geq 0$, which can be interpreted as the (generic and scaled) queue-length and workload processes, respectively. $x(t)$ is also an R-dimensional vector function of time $t \geq 0$, associated with the "free process" in the prelimit networks that captures the deviations of arrival and service processes from their means. The "residuals", $u(1)$ and $v(1)$, are R-dimensional vector constants.

The next lemma claims that the above can be approximated by the so-called fluid model specified in (3.11)–(3.15). And it is indeed a uniform continuity property if we consider all the processes involved in the $\mathcal{D}$-space (e.g., [1]) equipped with the uniform norm.

LEMMA B.1 (Uniform continuity). *Let T and M be any given positive numbers. For any ε, there exists a $\delta > 0$ such that for any $(q(t), w(t), x(t), u(1), v(1))$ satisfying* (B.1)–(B.5) *and*

$$|q(0)| + |u(1)| + |v(1)| \leq M, \qquad \sup_{0\leq t\leq T} |x(t)| < \delta, \tag{B.6}$$

we can find a fluid model $(\bar{q}(t), \bar{w}(t), \bar{u}(1), \bar{v}(1))$ satisfying (3.11)–(3.15) *and*

$$|\bar{q}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq M,$$

$$\sup_{0\leq t\leq T} |q(t) - \bar{q}(t)| + |u(1) - \bar{u}(1)| + |v(1) - \bar{v}(1)| < \varepsilon.$$

In addition, if the last part of the condition in (B.6) *is replaced by*

$$\sup_{0\leq t\leq T} |x(t)| + |u(1)| + |v(1)| < \delta,$$

we can further require $\bar{u}(1) = \bar{v}(1) = 0$ for the fluid model, which coincides with the standard and nondelayed version in (2.22)–(2.26).

PROOF. If to the contrary, we can find an $\varepsilon_0 > 0$ and a sequence $\{(q^{(i)}(t), w^{(i)}(t), x^{(i)}(t), u^{(i)}(1), v^{(i)}(1)); i = 1, 2, \ldots\}$ satisfying (B.1)–(B.5) and

$$|q^{(i)}(0)| + |u^{(i)}(1)| + |v^{(i)}(1)| \leq M, \qquad \lim_{i\to\infty} \sup_{0\leq t\leq T} |x^{(i)}(t)| = 0,$$

such that for all i, the following holds:

$$\sup_{0\leq t\leq T} |q^{(i)}(t) - \bar{q}(t)| + |u^{(i)}(1) - \bar{u}(1)| + |v^{(i)}(1) - \bar{v}(1)| \geq \varepsilon_0, \tag{B.7}$$

for any fluid model $(\bar{q}(t), \bar{w}(t), \bar{u}(1), \bar{v}(1))$ satisfying (3.11)–(3.15) with $|q(0)| + |u(1)| + |v(1)| \leq M$.

Applying the conventional approach for proving fluid limit theorem (refer to, e.g., [10, 15], among many others), however, we can find a subsequence of i such that as $i \to \infty$ along the subsequence, we have

$$\sup_{0\leq t\leq T} |q^{(i)}(t) - \bar{q}(t)| + |u^{(i)}(1) - \bar{u}(1)| + |v^{(i)}(1) - \bar{v}(1)| \to 0. \tag{B.8}$$

for some fluid model $(\bar{q}(t), \bar{u}(1), \bar{v}(1))$ satisfying (3.11)–(3.15) with $|q(0)| + |u(1)| + |v(1)| \leq M$, which contradicts to (B.7).

To claim the limit $(\bar{q}(t), \bar{u}(1), \bar{v}(1))$ in the above convergence as a solution to (3.11)–(3.15), it is necessary to verify (3.11)–(3.15) for the limit. Clearly, the technical details to do so are specific to the service discipline, but is routine given existing studies mentioned just now. Here, we illustrate those techniques in the context of static buffer priority discipline. Particularly, we must show that for the allocation process $\bar{d}(t)$ derived through the limit in (B.8), there is a set of nonnegative functions $\{\bar{\varsigma}_r(t)\}_{r\in\mathcal{R}}$ such that the specification in (2.27) is satisfied, that is,

$$\bar{d}_r(t) = \int_0^t \bar{\varsigma}_r(s)\,ds, \tag{B.9}$$

$$\sum_{r\in\mathcal{R}} \bar{\varsigma}_r(t) \leq 1 \quad \text{and} \quad \sum_{r\in H_r} \bar{\varsigma}_r(t) = 1 \qquad \text{if } \bar{q}_r(t) > 0. \tag{B.10}$$

First, note that the allocation processes $d^{(i)}(t)$, along with some nonnegative functions $\{\varsigma_r^{(i)}(t)\}_{r\in\mathcal{R}}$, also jointly satisfy the above relationships for each i. Then it is easy to see that there exists nonnegative functions $\{\bar{\varsigma}_r(t)\}_{r\in\mathcal{R}}$ such that the expression in (B.9) and the first condition in (B.10) hold. Second, suppose $\bar{q}_r(t) > 0$ for some fixed time $t \geq 0$ and the class r. Then we can find a number $\varepsilon > 0$ and a (small) interval $[t_1 t_2]$ satisfying $t_1 < t < t_2$ ($t_1 = 0$ if $t = 0$) such that for sufficiently large i we have for all $u \in [t_1 t_2]$, $q_r^{(i)}(u) > \varepsilon$. This implies for all $u \in [t_1 t_2]$, $\sum_{r\in H_r} \varsigma_r^{(i)}(u) = 1$ and, therefore, $\sum_{r\in H_r} d_r^{(i)}(u) = \sum_{r\in H_r} d_r^{(i)}(t_1) + (u - t_1)$. As $i \to \infty$, the later yields $\sum_{r\in H_r} \bar{d}_r(u) = \sum_{r\in H_r} \bar{d}_r(t_1) + (u - t_1)$, which implies the second condition in (B.10).

The additional part of the lemma is proved in the same manner. □

To apply the above lemma for proving Lemma 3.4, we first spell out the dynamics of the networks $\bar{Q}^{k,j}(u)$ in period $[0, T]$ (i.e., $\hat{Q}^k(m_k t)/y^k m_k$ in $[jy^k T/k, (j+1)y^k T/k]$ or $Q^k(t)/ky^k m_k$ in $[jky^k m_k T, (j+1)ky^k m_k T]$).

First, the original and unscaled arrival process, restarted at time $jky^k m_k T$, is also a (delayed) renewal process, which we denote as

$$E_r^{k,j}(t) = E_r^k\bigl(jky^k m_k T + t\bigr) - E_r^k\bigl(jky^k m_k T\bigr).$$

It is defined by the delayed starting time $U_r^{k,j}(0) := U_r^k(jky^k m_k T)$ (the "initial" residual arrival time) and the renewal sequence $\{u_r^k(i) : i \geq E_r^k(jky^k m_k T) + 2\}$. Denote the corresponding nondelayed version as $E_r^{o,k,j}(t)$, which is then defined by the renewal sequence $\{u_r^k(i) : i \geq E^k(jky^k m_k T) + 2\}$. Denote

$$\begin{aligned} \bar{U}_r^{k,j}(0) &= U_r^{k,j}(0)/ky^k m_k, \\ \bar{E}_r^{k,j}(u) &= E_r^{k,j}\bigl(ky^k m_k u\bigr)/ky^k m_k \quad \text{and} \\ \bar{E}_r^{o,k,j}(u) &= E_r^{o,k,j}\bigl(ky^k m_k u\bigr)/ky^k m_k. \end{aligned}$$

Then

$$\begin{aligned}\bar{E}_r^{k,j}(u)-\alpha_r^k u &= \bar{E}_r^{o,k,j}\big([u-\bar{U}_r^{k,j}(0)]^+\big)-\alpha_r^k[u-\bar{U}_r^{k,j}(0)]^+\\ &\quad +1_{\{u\geq\bar{U}_r^{k,j}(0)\}}/ky^k m_k-\alpha_r^k\big(u\wedge\bar{U}_r^{k,j}(0)\big).\end{aligned}$$

Second, the service process is characterized as

$$\begin{aligned}S_r^{k,j}(t)&=S_r^k\big(D_r^k(jky^k m_k T)+t\big)-S_r^k\big(D_r^k(jky^k m_k T)\big),\\ D_r^{k,j}(t)&=D_r^k\big(jky^k m_k T+t\big)-D_r^k\big(jky^k m_k T\big).\end{aligned}$$

That is, $S_r^{k,j}(t)$ is defined by the delayed starting time

$$V_r^{k,j}(0):=V_r^k\big(D_r^k(jky^k m_k T)\big)$$

(the "initial" residual arrival time) and the renewal sequence

$$\{v_r^k(i):i\geq S_r^k\big(D_r(jky^k m_k T)\big)+2\}.$$

Denote the corresponding nondelayed version as $S_r^{o,k,j}(t)$, which is then defined by the renewal sequence

$$\{v_r^k(i):i\geq S_r^k\big(D_r(jky^k m_k T)\big)+2\}.$$

Denote

$$\begin{aligned}\bar{V}_r^{k,j}(0)&=V_r^{k,j}(0)/ky^k m_k,\\ \bar{S}_r^{k,j}(u)&=S_r^{k,j}\big(ky^k m_k u\big)/ky^k m_k,\\ \bar{S}_r^{o,k,j}(u)&=S_r^{o,k,j}\big(ky^k m_k u\big)/ky^k m_k \quad\text{and}\\ \bar{D}_r^{k,j}(u)&=D_r^{k,j}\big(ky^k m_k u\big)/ky^k m_k.\end{aligned}$$

Then we write

$$\begin{aligned}\bar{S}_r^{k,j}(u)-\mu_r u &= \bar{S}_r^{o,k,j}\big([u-\bar{V}_r^{k,j}(0)]^+\big)-\mu_r[u-\bar{V}_r^{k,j}(0)]^+\\ &\quad +1_{\{u\geq\bar{V}_r^{k,j}(0)\}}/ky^k m_k-\mu_r\big(u\wedge\bar{V}_r^{k,j}(0)\big).\end{aligned}\tag{B.11}$$

Note that $\bar{D}_r^{k,j}(u)$, the allocation process, must satisfy conditions specific to the service discipline, like the one in (2.8), or the one in (2.27) for the corresponding fluid model below. (In the discussion below, for ease of exposition, we will not describe and verify such conditions further, which is routine, though tedious, given the discussions above and the extensive related literatures.)

Third, the routing process and its fluid scaling are denoted as, for $s, r \in \mathcal{R}$,

$$\begin{aligned}
\Phi^{k,j}_{sr}(n) &= \Phi^{k}_{sr}\big(S^{k}_{s}\big(D^{k}_{s}(jky^{k}m_{k}T)\big) + n\big) - \Phi^{k}_{sr}\big(S^{k}_{r}\big(D_{r}(jky^{k}m_{k}T)\big)\big),\\
\bar{\Phi}^{k,j}_{sr}(u) &= \frac{1}{ky^{k}m_{k}}\Phi^{k,j}_{sr}\big(\lfloor ky^{k}m_{k}u\rfloor\big).
\end{aligned}$$

Finally, the queue-length process can be written as

$$\begin{aligned}
\bar{Q}^{k,j}_{r}(u) &= \bar{Q}^{k,j}_{r}(0) + \bar{E}^{k,j}_{r}(u)\\
&\quad + \sum_{s} \bar{\Phi}^{k,j}_{sr}\big(\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big)\big) - \bar{S}^{k,j}_{r}\big(\bar{D}^{k,j}_{r}(u)\big)\\
&= \bar{Q}^{k,j}_{r}(0) + \big(\bar{E}^{k,j}_{r}(u) - \alpha^{k}_{r}u\big) - \big(\bar{S}^{k,j}_{r}\big(\bar{D}^{k,j}_{r}(u)\big) - \mu_{r}\bar{D}^{k,j}_{r}(u)\big)\\
&\quad + \sum_{s} p_{sr}\big(\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big) - \mu_{s}\bar{D}^{k,j}_{s}(u)\big)\\
&\quad + \sum_{s}\big(\bar{\Phi}^{k,j}_{sr}\big(\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big)\big) - p_{sr}\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big)\big)\\
&\quad + \big(\alpha^{k}_{r}u - \alpha_{r}u\big) + \Big(\alpha_{r}u - \mu_{r}\bar{D}^{k,j}_{r}(u) + \sum_{s} p_{sr}\mu_{s}\bar{D}^{k,j}_{s}(u)\Big)\\
&= \bar{Q}^{k,j}(0) - \alpha_{r}\big(u \wedge \bar{U}^{k,j}_{r}(0)\big)\\
&\quad + \mu_{r}\big(\bar{D}^{k,j}_{r}(u) \wedge \bar{V}^{k,j}_{r}(0)\big) + \sum_{s} p_{sr}\mu_{s}\big(\bar{D}^{k,j}_{s}(u) \wedge \bar{V}^{k,j}_{s}(0)\big)\\
&\quad + \big(\bar{E}^{o,k,j}_{r}\big([u - \bar{U}^{k,j}_{r}(0)]^{+}\big) - \alpha^{k}_{r}[u - \bar{U}^{k,j}_{r}(0)]^{+}\big)\\
&\quad - \big(\bar{S}^{o,k,j}_{r}\big([\bar{D}^{k,j}_{r}(u) - \bar{V}^{k,j}_{r}(0)]^{+}\big) - \mu_{r}[\bar{D}^{k,j}_{r}(u) - \bar{V}^{k,j}_{r}(0)]^{+}\big)\\
&\quad + \sum_{s} p_{sr}\big(\bar{S}^{o,k,j}_{s}\big([\bar{D}^{k,j}_{s}(u) - \bar{V}^{k,j}_{s}(0)]^{+}\big)\\
&\quad - \mu_{s}[\bar{D}^{k,j}_{s}(u) - \bar{V}^{k,j}_{s}(0)]^{+}\big)\\
&\quad + \sum_{s}\big(\bar{\Phi}^{k,j}_{sr}\big(\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big)\big) - p_{sr}\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big)\big)\\
&\quad + \frac{1}{ky^{k}m_{k}}\Big(1_{\{u \ge \bar{U}^{k,j}_{r}(0)\}} - 1_{\{\bar{D}^{k,j}_{r}(u) \ge \bar{V}^{k,j}_{r}(0)\}}\\
&\quad + \sum_{s} p_{sr}1_{\{\bar{D}^{k,j}_{s}(u) \ge \bar{V}^{k,j}_{s}(0)\}}\Big)\\
&\quad - \big(\alpha^{k}_{r} - \alpha_{r}\big)\big(u \wedge \bar{U}^{k,j}_{r}(0)\big) + \big(\alpha^{k}_{r} - \alpha_{r}\big)u\\
&\quad + \Big(\alpha_{r}u - \mu_{r}\bar{D}^{k,j}_{r}(u) + \sum_{s} p_{sr}\mu_{s}\bar{D}^{k,j}_{s}(u)\Big).
\end{aligned}\tag{B.12}$$

PROOF OF LEMMA 3.4. To apply Lemma B.1, we denote the terms in (B.12) which we show will vanish as follows for convenience:

$$
\begin{aligned}
x_r(u) :={}& \Sigma_E - \Sigma_S + \Sigma_S^+ + \Sigma_\Phi + (\alpha_r^k - \alpha_r)(u - u \wedge \bar{U}_r^{k,j}(0)) \\
&+ \frac{1}{ky^k m_k}\Big(1_{\{u \geq \bar{U}_r^{k,j}(0)\}} - 1_{\{\bar{D}_r^{k,j}(u) \geq \bar{V}_r^{k,j}(0)\}} \\
&+ \sum_s p_{sr} 1_{\{\bar{D}_s^{k,j}(u) \geq \bar{V}_s^{k,j}(0)\}}\Big),
\end{aligned}
\tag{B.13}
$$

where

$$
\begin{aligned}
\Sigma_E &:= \bar{E}_r^{o,k,j}([u - \bar{U}_r^{k,j}(0)]^+) - \alpha_r^k[u - \bar{U}_r^{k,j}(0)]^+, \\
\Sigma_S &:= \bar{S}_r^{o,k,j}([\bar{D}_r^{k,j}(u) - \bar{V}_r^{k,j}(0)]^+) - \mu_r[\bar{D}_r^{k,j}(u) - \bar{V}_r^{k,j}(0)]^+, \\
\Sigma_S^+ &:= \sum_s p_{sr}(\bar{S}_s^{o,k,j}([\bar{D}_s^{k,j}(u) - \bar{V}_s^{k,j}(0)]^+) - \mu_s[\bar{D}_s^{k,j}(u) - \bar{V}_s^{k,j}(0)]^+), \\
\Sigma_\Phi &:= \sum_s (\bar{\Phi}_{sr}^{k,j}(\bar{S}_s^{k,j}(\bar{D}_s^{k,j}(u))) - p_{sr}\bar{S}_s^{k,j}(\bar{D}_s^{k,j}(u))).
\end{aligned}
$$

First, we estimate the term involving the arrival process, Σ_E. For any $u \in [0, T]$, we have

$$
\begin{aligned}
|\Sigma_E| &= \frac{1}{ky^k m_k}|E_r^{o,k}(km_k\tau + ky^k m_k[u - \bar{U}_r^{k,j}(0)]^+) \\
&\quad - E_r^{o,k}(km_k\tau) - ky^k m_k \alpha_r^k[u - \bar{U}_r^{k,j}(0)]^+| \\
&\leq \sup_{u' \in [0,T]} \frac{1}{y^k}\left|\frac{1}{km_k}(E_r^{o,k}(km_k(\tau + y^k u'))\right. \\
&\quad \left. - E_r^{o,k}(km_k\tau)) - ky^k\alpha_r^k u'\right| \\
&= \frac{1}{y^k}\sup_{u' \in [0,T]}\left|\frac{1}{m_k}(\bar{E}_r^{o,k}(m_k(\tau + y^k u')) - \bar{E}_r^{o,k}(m_k\tau)) - y^k\alpha_r^k u'\right| \\
&\leq \frac{1}{y^k}\sum_{i=1}^{\lceil y^k \rceil}\sup_{u' \in [0,T]}\left|\frac{1}{m_k}(\bar{E}_r^{o,k}(m_k(\tau + (i-1)T + u'))\right. \\
&\quad \left. - \bar{E}_r^{o,k}(m_k(\tau + (i-1)T))) - \alpha_r^k u'\right|.
\end{aligned}
\tag{B.14}
$$

[Here, we denote $\tau := jy^kT - U_r^{k,0}(0)/km_k + U_r^{k,j}(0)/km_k$ for convenience.] Estimate the time $\tau + (i-1)T$ inside the supremum above for $j < k\Delta/y^kT$,

$i \le \lceil y^k \rceil$,

$$\begin{aligned}\tau + (i-1)T &\le j y^k T - \frac{U_r^{k,0}(0)}{km_k} + \frac{U_r^{k,j}(0)}{km_k} + y^k T\\ &\le \frac{k\Delta}{y^k T} y^k T + y^k \bar{U}_r^{k,j}(0) + y^k T \le k\Delta + y^k M + y^k T.\end{aligned}$$

Since $|\hat{W}^k(0)/m_k| \le 1$, we have $y^k \le 1 + k/\log k$ for $\omega \in \Omega_X^k(\bar{\Delta}, m_k)$. The above estimate implies

$$\tau + (i-1)T \le k\big(\Delta + O(1/\log k)\big) \le k\bar{\Delta}$$

for sufficiently large k. The above inequality indicates that the time periods involved in (B.14) fall within those covered in $\Omega_E^k(\bar{\Delta}, T, m_k)$, so that the bound in that event can be invoked for each item in (B.14); that is, we have for sufficiently large k, for any $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, for all $i = 1, \dots, \lceil y^k \rceil$,

$$\begin{aligned}&\sup_{u' \in [0,T]} \bigg| \frac{1}{m_k} \big(\bar{E}_r^{o,k}\big(m_k\big(\tau + (i-1)T + u'\big)\big)\\ &\qquad - \bar{E}_r^{o,k}\big(m_k\big(\tau + (i-1)T\big)\big)\big) - \alpha_r^k u' \bigg|\\ &\qquad \le \sup_{t \in [0, k\bar{\Delta}]} \sup_{u' \in [0, \bar{T}]} \bigg| \frac{1}{m_k} \big(\bar{E}_r^{o,k}\big(m_k(t + u')\big) - \bar{E}_r^{o,k}(m_k t)\big) - \alpha_r^k u' \bigg|\\ &\qquad \le \delta,\end{aligned} \tag{B.15}$$

where the second inequality follows from the definition of $\Omega^k(\bar{\Delta}, T, m_k)$ ($\subset \Omega_E^k(\bar{\Delta}, T, m_k)$) with sufficiently large k (say, $1/\log k < \delta$). Then we have from (B.14) and (B.15),

$$\Sigma_E \le \frac{1}{y^k} \lceil y^k \rceil \delta \le 2\delta. \tag{B.16}$$

Second, we estimate the terms involving the service process, Σ_S and Σ_S^+. The approach is similar to the estimation of Σ_E. For any $u \in [0, T]$, we have

$$\begin{aligned}|\Sigma_S| &= \frac{1}{ky^k m_k} \big| S_r^{o,k}\big(km_k\tau + ky^k m_k\big[\bar{D}_r^{k,j}(u) - \bar{V}_r^{k,j}(0)\big]^+\big) - S_r^{o,k}(km_k\tau)\\ &\qquad - ky^k m_k \mu_r \big[\bar{D}_r^{k,j}(u) - \bar{V}_r^{k,j}(0)\big]^+ \big|\\ &\le \sup_{u' \in [0,T]} \frac{1}{ky^k m_k} \big| S_r^{o,k}\big(km_k\tau + ky^k m_k u'\big) - S_r^{o,k}(km_k\tau) - ky^k m_k \mu_r u' \big|\\ &= \frac{1}{y^k} \sup_{u' \in [0,T]} \bigg| \frac{1}{m_k} \big(\bar{S}_r^{o,k}\big(m_k(\tau + y^k u')\big) - \bar{S}_r^{o,k}(m_k\tau)\big) - y^k \mu_r u' \bigg|\end{aligned}$$

$$\begin{aligned}&\le \frac{1}{y^k}\sum_{i=1}^{\lceil y^k\rceil}\sup_{u'\in[0,T]}\biggl|\frac{1}{m_k}\bigl(\bar{S}_r^{o,k}\bigl(m_k\bigl(\tau+(i-1)T+u'\bigr)\bigr)\\&\qquad - \bar{S}_r^{o,k}\bigl(m_k\bigl(\tau+(i-1)T\bigr)\bigr)\bigr)-\mu_r u'\biggr|.\end{aligned}$$

[Here, reusing the notation, we denote $\tau := D_r^k(jky^k m_k T)/km_k - U_r^{k,0}(0)/km_k + U_r^{k,j}(0)/km_k$ for convenience.] The first inequality in the above is because for $u \in [0, T]$,

$$\bigl[\bar{D}_r^{k,j}(u) - \bar{V}_r^{k,j}(0)\bigr]^+ \le \bar{D}_r^{k,j}(u) \le T.$$

Estimate the time $\tau + (i-1)T$ inside the supremum above for $j < k\Delta/y^k T$, $i < \lceil y^k\rceil$,

$$\begin{aligned}\tau+(i-1)T &\le \frac{D_r^k(jky^k m_k T)}{km_k} - \frac{U_r^{k,0}(0)}{km_k} + \frac{U_r^{k,j}(0)}{km_k} + y^k T\\&\le y^k jT + y^k \bar{U}_r^{k,j}(0) + y^k T \le k\Delta + y^k M + y^k T\\&\le k\bigl(\Delta + O(1/\log k)\bigr) \le k\bar{\Delta},\end{aligned}$$

for $\omega \in \Omega_X^k(\bar{\Delta}, m_k)$ and sufficiently large k. Hence, we have for sufficiently large k, for any $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$,

$$\begin{aligned}&\sup_{u'\in[0,T]}\biggl|\frac{1}{m_k}\bigl(\bar{S}_r^{o,k}\bigl(m_k\bigl(\tau+(i-1)T+u'\bigr)\bigr) - \bar{S}_r^{o,k}\bigl(m_k\bigl(\tau+(i-1)T\bigr)\bigr)\bigr)-\mu_r u'\biggr|\\&\quad\le \sup_{t\in[0,k\bar{\Delta}]}\sup_{u'\in[0,T]}\biggl|\frac{1}{m_k}\bigl(\bar{S}_r^{o,k}\bigl(m_k(t+u')\bigr) - \bar{S}_r^{o,k}(m_k t)\bigr)-\mu_r u'\biggr| \le \delta,\end{aligned}$$

and thereafter,

$$\Sigma_S \le 2\delta. \tag{B.17}$$

This immediately yields the following for some constant κ':

$$\Sigma_S^+ \le \kappa'\delta. \tag{B.18}$$

Third, we estimate the term involving the routing process, Σ_Φ. Note that $\bar{D}_r^{k,j}(u) \le u$ by definition. Then, from (B.11), we have for sufficiently large k, for any $m_k \ge |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$, $u \in [0, T]$,

$$\begin{aligned}\bar{S}_r^{k,j}\bigl(\bar{D}_r^{k,j}(u)\bigr) &= \mu_r \bar{D}_r^{k,j}(u) + \Sigma_S + 1_{\{\bar{D}_r^{k,j}(u)\ge \bar{V}_r^{k,j}(0)\}}/ky^k m_k\\&\quad - \mu_r\bigl(\bar{D}_r^{k,j}(u) \wedge \bar{V}_r^{k,j}(0)\bigr).\\&\le \mu_r u + 2\delta + 1 \le 2\mu_r T \le 2\kappa_\mu T.\end{aligned}$$

Using the above bound, we estimate each term in Σ_Φ, denoted as $\Sigma_{\Phi,s}$:

$$\begin{aligned}
|\Sigma_{\Phi,s}| &= \big|\bar{\Phi}^{k,j}_{sr}\big(\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big)\big) - p_{sr}\bar{S}^{k,j}_{s}\big(\bar{D}^{k,j}_{s}(u)\big)\big| \\
&\le \sup_{u'\in[0,\kappa_\mu T]} \big|\bar{\Phi}^{k,j}_{sr}(u') - p_{sr}u'\big| \\
&= \sup_{u'\in[0,\kappa_\mu T]} \bigg|\frac{1}{ky^k m_k}\big(\Phi^k_{sr}\big(S^k_s\big(D^k_s(jky^k m_k T)\big) + \lfloor ky^k m_k u'\rfloor\big) \\
&\qquad - \Phi^k_{sr}\big(S^k_r\big(D_r(jky^k m_k T)\big)\big)\big) - p_{sr}u'\bigg| \\
&= \frac{1}{y^k}\sup_{u'\in[0,\kappa_\mu T]} \bigg|\frac{1}{m_k}\big(\bar{\Phi}^k_{sr}\big(\bar{S}^k_s\big(\bar{D}^k_s(jy^k m_k T)\big) + y^k m_k u'\big) \\
&\qquad - \bar{\Phi}^k_{sr}\big(\bar{S}^k_r\big(\bar{D}_r(jy^k m_k T)\big)\big)\big) - p_{sr}y^k u'\bigg| \\
&= \frac{1}{y^k}\sup_{u'\in[0,\kappa_\mu T]} \bigg|\frac{1}{m_k}\big(\bar{\Phi}^k_{sr}\big(m_k(\tau + y^k u')\big) - \bar{\Phi}^k_{sr}(m_k\tau)\big) - p_{sr}y^k u'\bigg| \\
&\le \frac{1}{y^k}\sum_{i=1}^{\lceil y^k\rceil}\sup_{u'\in[0,\kappa_\mu T]} \bigg|\frac{1}{m_k}\big(\bar{\Phi}^k_{sr}\big(m_k(\tau + (i-1)\kappa_\mu T + u')\big) \\
&\qquad - \bar{\Phi}^k_{sr}\big(m_k(\tau + (i-1)\kappa_\mu T)\big)\big) - p_{sr}u'\bigg|.
\end{aligned}$$

[Here, reusing the notation again, we denote $\tau := \bar{S}^k_s(\bar{D}^k_s(jy^k m_k T))/m_k$ for convenience.] Estimate the time $\tau + (i-1)\kappa_\mu T$ inside the supremum above for $j < k\Delta/y^k T, i < \lceil y^k \rceil$,

$$\begin{aligned}
\tau + (i-1)\kappa_\mu T &\le \frac{1}{m_k}\bar{S}^k_s(jy^k m_k T) + y^k\kappa_\mu T \\
&\le \frac{1}{m_k}\bar{S}^k_s(m_k\cdot k\Delta) + \left(1 + \frac{k}{\log k}\right)\kappa_\mu T \\
&\le \frac{1}{m_k}\bar{S}^{o,k}_s(m_k\cdot k\Delta) + 1 + \left(1 + \frac{k}{\log k}\right)\kappa_\mu T \\
&\le \left(\mu_s\cdot k\Delta + \frac{k\Delta}{\kappa_\mu T}\cdot\frac{1}{\log k}\right) + 1 + \left(1 + \frac{k}{\log k}\right)\kappa_\mu T, \\
&\le k\kappa_\mu(\Delta + O(1/\log k) \le k\kappa_\mu\bar{\Delta},
\end{aligned}$$

for $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$ and sufficiently large k. [Note that the fourth inequality above is due to the bound in the definition of $\Omega^k_S(\bar{\Delta}, T, m_k)$.] Hence, we have for sufficiently large k, for any $m_k \geq |\hat{\Xi}^k(0)| \vee 1$ and $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$,

$$\begin{aligned}
&\sup_{u' \in [0,\kappa_\mu T]} \biggl| \frac{1}{m_k} (\bar{\Phi}^k_{sr}(m_k(\tau + (i-1)\kappa_\mu T + u')) \\
&\quad - \bar{\Phi}^k_{sr}(m_k(\tau + (i-1)\kappa_\mu T))) - p_{sr} u' \biggr| \\
&\quad \leq \sup_{t \in [0,k\kappa_\mu \bar{\Delta}]} \sup_{u' \in [0,\kappa_\mu T]} \biggl| \frac{1}{m_k} (\Phi^k_{sr}(m_k(t+u')) - \Phi^k_{sr}(m_k t)) - p_{sr} u' \biggr| \\
&\quad \leq \delta,
\end{aligned}$$

and thereafter,

$$\Sigma_{\Phi,s} \leq 2\delta.$$

This yields the following for some constant κ'' immediately:

$$\Sigma_\Phi \leq \kappa'' \delta. \tag{B.19}$$

From (B.13), (B.16), (B.17), (B.18), (B.19), we know that the condition in (B.6) in Lemma B.1 [in particular, $\sup_{0 \leq t \leq T} |x(t)| < \delta$, with δ redefined properly] can be justified for all sufficiently large k, all $j = 0, 1, \dots, k\Delta/y^k T$ and all $\omega \in \Omega^k(\bar{\Delta}, T, m_k)$. Moreover, verifying the other conditions in (B.1)–(B.5) is rather straightforward. Therefore, Lemma B.1 can be invoked to claim the conclusion in (a).

We sketch the proof for part (b) only. Note that the time T can be chosen such that

$$T > \bar{U}^{k,0}_r(0), \qquad \bar{D}^{k,0}_r(T) > \bar{V}^{k,0}_r(0), \tag{B.20}$$

for $r \in \mathcal{R}$, for sufficiently large k. The second inequality is a consequence of the conclusion in (a) and the uniform attraction property in Theorem 3.3: given the bounded initial state $|\bar{\Xi}^{k,0}(0)| = |\hat{\Xi}^k(0)|/y^k m_k \leq 1$, the process $\bar{D}^{k,0}_r(t)$ is close to $\bar{d}_r(t)$ for sufficiently large k, whereas $\bar{d}_r(t)$ is close to $(\lambda_r/\mu_r)t$ for sufficiently large t. The inequalities in (B.20) implies that for $j \geq 1$, $U^{k,j}_r(0)$ [resp., $V^{k,j}_r(0)$] must be a portion of an interarrival time (resp., a service time) of class-r *other than* the initial residual arrival time $u^k_r(1)$ [resp., the initial service time $v^k_r(1)$] of the original kth network. Hence, following the definition in (3.1), (3.2), we have for $1 \leq j \leq k\bar{\Delta}/y^k T$ and $r \in \mathcal{R}$,

$$U^{k,j}_r(0) \leq u^{k,\max}_r(k^2 m_k \bar{\Delta}), \qquad V^{k,j}_r(0) \leq v^{k,\max}_r(k^2 m_k \bar{\Delta}).$$

Consequently, give the assumption $\omega \in \Omega^k(\bar{\Delta}, T, m_k) \subset \Omega^k_u(\bar{\Delta}, m_k) \cap \Omega^k_v(\bar{\Delta}, m_k)$, the above inequalities imply the first conclusion in part (b), which along with the last conclusion in Lemma B.1, further implies the second conclusion in part (b). □

REFERENCES

[1] BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed. Wiley, New York. MR1700749

[2] BRAMSON, M. (1996). Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Syst. Theory Appl.* **22** 5–45. MR1393404

[3] BRAMSON, M. (1996). Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Syst. Theory Appl.* **23** 1–26. MR1433762

[4] BRAMSON, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst. Theory Appl.* **30** 89–148. MR1663763

[5] BRAMSON, M. (1998). Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Syst. Theory Appl.* **28** 7–31. MR1628469

[6] BRAMSON, M. and DAI, J. G. (2001). Heavy traffic limits for some queueing networks. *Ann. Appl. Probab.* **11** 49–90. MR1825460

[7] BRAVERMAN, A. and DAI, J. G. (2017). Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems. *Ann. Appl. Probab.* **27** 550–581. MR3619795

[8] BRAVERMAN, A., DAI, J. G. and FENG, J. (2016). Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.* **6** 301–366. MR3633538

[9] BUDHIRAJA, A. and LEE, C. (2009). Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.* **34** 45–56. MR2542988

[10] CHEN, H. (1995). Fluid approximations and stability of multiclass queueing networks: Work-conserving disciplines. *Ann. Appl. Probab.* **5** 637–665. MR1359823

[11] CHEN, H. and YE, H. Q. (2012). Asymptotic optimality of balanced routing. *Oper. Res.* **60** 163–179. MR2911665

[12] CHEN, H. and YE, H. Q. (2001). Existence condition for the diffusion approximations of multiclass priority queueing networks. *Queueing Syst.* **38** 435–470. MR1856548

[13] CHEN, H. and ZHANG, H. (2000). Diffusion approximations for some multiclass queueing networks with FIFO service disciplines. *Math. Oper. Res.* **25** 679–707. MR1855373

[14] CHEN, H. and ZHANG, H. (2000). A sufficient condition and a necessary condition for the diffusion approximations of multiclass queueing networks under priority service disciplines. *Queueing Syst. Theory Appl.* **34** 237–268. MR1769773

[15] DAI, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77. MR1325041

[16] DAI, J. G., DIEKER, A. B. and GAO, X. (2014). Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Syst.* **78** 1–29. MR3238006

[17] DAI, J. G. and HARRISON, J. M. (1992). Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis. *Ann. Appl. Probab.* **2** 65–86. MR1143393

[18] DAI, J. G. and LIN, W. (2008). Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.* **18** 2239–2299. MR2473656

[19] DAI, J. G. and MEYN, S. P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Automat. Control* **40** 1889–1904. MR1358006

[20] DAVIS, M. H. A. (1984). Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *J. Roy. Statist. Soc. Ser. B* **46** 353–388. MR0790622

[21] DUPUIS, P. and WILLIAMS, R. J. (1994). Lyapunov functions for semimartingale reflecting Brownian motions. *Ann. Probab.* **22** 680–702. MR1288127

[22] DURRETT, R. (2010). *Probability: Theory and Examples*, 4th ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **31**. Cambridge Univ. Press, Cambridge. MR2722836

[23] GAMARNIK, D. and GOLDBERG, D. A. (2013). On the rate of convergence to stationarity of the M/M/N queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* **23** 1879–1912. MR3134725

[24] GAMARNIK, D. and STOLYAR, A. L. (2012). Multiclass multiserver queueing system in the Halfin–Whitt heavy traffic regime: Asymptotics of the stationary distribution. *Queueing Syst.* **71** 25–51. MR2925789

[25] GAMARNIK, D. and ZEEVI, A. (2006). Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Ann. Appl. Probab.* **16** 56–90. MR2209336

[26] GURVICH, I. (2014). Validity of heavy-traffic steady-state approximations in multiclass queueing networks: The case of queue-ratio disciplines. *Math. Oper. Res.* **39** 121–162. MR3173006

[27] GUT, A. (1988). *Stopped Random Walks*: *Limit Theorems and Applications*. *Applied Probability. A Series of the Applied Probability Trust* **5**. Springer, New York. MR0916870

[28] KANG, W. N., KELLY, F. P., LEE, N. H. and WILLIAMS, R. J. (2009). State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probab.* **19** 1719–1780. MR2569806

[29] KATSUDA, T. (2010). State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic. *Queueing Syst.* **65** 237–273. MR2652044

[30] KATSUDA, T. (2012). Stationary distribution convergence for a multiclass single-server queue in heavy traffic. *Sci. Math. Jpn.* **75** 317–334. MR3099762

[31] KRICHAGINA, E. V. and TAKSAR, M. I. (1992). Diffusion approximation for GI/G/1 controlled queues. *Queueing Syst. Theory Appl.* **12** 333–367. MR1200872

[32] MANDELBAUM, A. and STOLYAR, A. L. (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52** 836–855. MR2104141

[33] RYBKO, A. N. and STOLYAR, A. L. (1992). On the ergodicity of random processes that describe the functioning of open queueing networks. *Problemy Peredachi Informatsii* **28** 3–26. MR1189331

[34] SHAH, D., TSITSIKLIS, J. N. and ZHONG, Y. (2014). Qualitative properties of α-fair policies in bandwidth-sharing networks. *Ann. Appl. Probab.* **24** 76–113. MR3161642

[35] STOLYAR, A. L. (2004). Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.* **14** 1–53. MR2023015

[36] STOLYAR, A. L. (2015). Diffusion-scale tightness of invariant distributions of a large-scale flexible service system. *Adv. in Appl. Probab.* **47** 251–269. MR3327324

[37] STOLYAR, A. L. (2015). Tightness of stationary distributions of a flexible-server system in the Halfin–Whitt asymptotic regime. *Stoch. Syst.* **5** 239–267. MR3442428

[38] STOLYAR, A. L. and YUDOVINA, E. (2012). Tightness of invariant distributions of a large-scale flexible service system under a priority discipline. *Stoch. Syst.* **2** 381–408. MR3354771

[39] WILLIAMS, R. J. (1998). Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Syst. Theory Appl.* **30** 27–88. MR1663759

[40] YE, H. Q. and YAO, D. D. (2008). Heavy-traffic optimality of a stochastic network under utility-maximizing resource allocation. *Oper. Res.* **56** 453–470. MR2410316

[41] YE, H. Q. and YAO, D. D. (2010). Utility-maximizing resource control: Diffusion limit and asymptotic optimality for a two-bottleneck model. *Oper. Res.* **58** 613–623. MR2680567

[42] YE, H. Q. and YAO, D. D. (2012). A stochastic network under proportional fair resource control-diffusion limit with multiple bottlenecks. *Oper. Res.* **60** 716–738. MR2960540

[43] YE, H. Q. and YAO, D. D. (2016). Diffusion limit of fair resource control—Stationarity and interchange of limits. *Math. Oper. Res.* **41** 1161–1207. MR3544792

DEPARTMENT OF LOGISTICS
AND MARITIME STUDIES
HONG KONG POLYTECHNIC UNIVERSITY
HUNG HOM, KOWLOON
HONG KONG
E-MAIL: lgtyehq@polyu.edu.hk

DEPARTMENT OF INDUSTRIAL ENGINEERING
AND OPERATIONS RESEARCH
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: yao@ieor.columbia.edu