MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Diffusion Approximation for Fair Resource Control — Interchange of Limits under a Moment Condition

Heng-Qing Ye
Dept of Logistics and Maritime Studies, Hong Kong Polytechnic University, Hong Kong, lgtyehq@polyu.edu.hk,

David D. Yao
Dept of Industrial Engineering and Operations Research, Columbia University, New York, USA, yao@columbia.edu

In a prior study [20], focusing on a class of stochastic processing network with fair resource control, we have justified the diffusion approximation (in the context of the interchange of limits) provided that the p-th moment of the workloads are bounded. To this end, we have introduced the so-called bounded workload condition that requires the workload process be bounded by a free process plus the initial workload. This condition is for a derived process, the workload, as opposed to primitives such as arrival processes and service requirements; as such, it could be difficult to verify. In this paper we establish the interchange of limits under a moment condition of suitable order on the *primitives* directly: the required order is $p^* > 2(p+2)$ on the moments of the primitive processes so as to bound the p-th moment of the workload. This moment condition is trivial to verify, and indeed automatically holds in networks where the primitives have moments of all orders, for instance, renewal arrivals with phase-type interarrival times and i.i.d. phase-type service times.

Keywords: resource-sharing network, diffusion limit, stationary distribution, interchange of limits, uniform stability.

---

1. Introduction This is a follow-up study on our recent work [20], where we justify the interchange of limits in a stochastic processing network under fair resource control. This is a multiclass queueing network model with the additional features that (a) each resource (server) is shared among the job classes according to a specific "proportional fair" mechanism, and (b) to be processed in the network each job class may require the simultaneous occupancy of more than one resource. To evaluate the steady-state performance of such a network, the so-called diffusion approximation (or heavy-traffic steady-state approximation) appears to be the only analytically viable alternative to simulation. The idea is to scale, in both time and space, the stochastic processes of interest (e.g., those associated with queue lengths or workloads) in the original network, and to use their limits, which are often characterized by diffusion processes, as approximations for the performance of the original network.

To illustrate this idea more precisely, we follow the formalism in Gamarnik and Zeevi [9] to use the rectangle in Figure 1. Let $W(t)$, a vector process, denote the workload at time t in the original network. Consider an infinite sequence of copies (or variations) of the original network, indexed by k. Let $W^k(t)$ denote the workload associated with the k-th network in the sequence; and let $\hat{W}^k(t) := W^k(k^2t)/k$ denote its diffusion-scaled version. Then, following edges I and III in the rectangle, i.e., taking $k \to \infty$ and then $t \to \infty$, we will reach the diffusion limit, $\hat{W}(\infty)$, and use it as an approximation for the steady-state workload in the original network. Yet, the original network is represented by the k-th network in the sequence, for some k large enough; hence, its steady-state workload is best approximated by $\lim_k \hat{W}^k(\infty)$. This last limit corresponds to taking $t \to \infty$ and then $k \to \infty$, i.e., following edges II and IV in the rectangle.

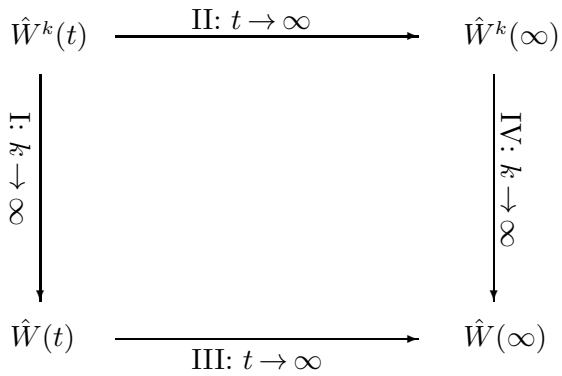

FIGURE 1. Interchange of limits

Thus, to justify the diffusion approximation is tantamount to verifying the following interchange of limits:

$$\lim_{t\to\infty}\lim_{k\to\infty}\hat{W}^k(t) = \lim_{k\to\infty}\lim_{t\to\infty}\hat{W}^k(t). \tag{1}$$

Central to this justification is of course to identify the "right" conditions — as general as possible, easy to verify, etc — under which the above equation (in some specific sense) can be proven. This type of justifications has in recent years been carried out for select queueing network models, and a brief overview of the related literature is in order. (Refer to [20], §1.1 and §1.2, for a more detailed review.)

Related Literature Pioneering studies on the interchange of limits to justify the diffusion approximation for the generalized Jackson network, a single-class queueing network include Gamarnik and Zeevi [9], and Budhiraja and Lee [3]. In the multiclass setting, Shah *et al* [16] proves the interchange of limits in a resource-sharing network with Poisson arrivals and i.i.d. exponential service times, and thus exploits certain Markovian and martingale properties in this setting. The same Markovian model (but allowing the service times to follow a certain class of phase-type distributions) has been studied more recently in Wang *et al* [17], where explicit upper- and lower-bounds are developed for the expected (weighted) sum of jobs in the network under steady state. Under heavy-traffic scaling, the bounds are insensitive to the service-time distributions; thus, the result also provides a justification for the interchange of limits. In another study, Gurvich [10] examines the interchange of limits for a class of multiclass queueing networks under a condition that requires the fluid model associated with the sequence of networks converges under heavy traffic to the fixed-point state space at a *linear rate*.

In our own prior work [20], we have justified the interchange with a bounded workload condition — that the workload can be bounded by a "free process" plus the initial workload. However, this condition is for a derived process, the workload, as opposed to primitives such as arrival processes and service requirements. Thus, verifying this condition could be highly non-trivial (as illustrated by the several examples in [20]). To overcome this difficulty, in this paper we show the interchange of limits in (1) can be accomplished by requiring a moment condition of suitable order on the primitive processes. Specifically, the required order is $p^* > 2(p+2)$ on the moments of the primitive processes in order to bound the p-th moment of the workload. This condition is trivial to verify, and indeed automatically holds in networks where the primitives have moments of all orders, for instance, renewal arrivals with phase-type interarrival times and i.i.d. phase-type service times. (Note, however, this p^* moment condition is not more general—i.e., weaker—than the bounded workload condition, since the latter leads to the interchange without having to assume higher-order moments on the primitives.)

Contributions and Organization It turns out that to justify the interchange of limits under the p^*-th moment condition requires overcoming several serious technical difficulties that are not present in [20]. Specifically, we need to focus on a sequence of "regular" events, under which the relevant processes behave "nicely," and the probabilities of these events occurring will approach 1 at a certain rate (cf. Lemma 3). We can then apply Bramson's [2] "hydro-dynamic" approach to show that the bounded workload condition holds for sample paths in the regular events (cf. Proposition 4); hence, the workload process, when restricted to the regular events, possesses a bounded p-th moment. Furthermore, the p-th moment of the workload process restricted to the small-probability, non-regular events, follows the same bound. Combining these two cases leads to the desired result.

We have successfully applied the same p^*-th moment condition in another recent study [21] to justify the interchange of limits in traditional multiclass queueing networks (MQN). In comparison, the resource-sharing network (RSN) studied here has at least two distinct features, concurrent resource occupancy and real-time resource allocation, that are not present in MQN. These features bring forth new technical issues in identifying the regular events (alluded to above) and bounding the workload associated such events, and in establishing the uniform stability and the uniform continuity; refer to Figure 2 below. Consider, for instance, the complementarity property, which is central to establishing the uniform bound for workloads in regular events (Proposition 4). While in MQN the complementarity property holds for all pre-limit networks, as well as the limit, this is not the case in RSN thanks to the two features mentioned above; and to overcome this difficulty is a major technical challenge.

Figure 2 below is a high-level illustration of the logical relations among the various technical results presented in this paper. In particular, results circled inside the dotted rectangle constitute the main contributions of this paper; whereas results from the companion paper [20], along with exactly what roles they play, are marked out in boxes on the right half of the figure, including several lemmas (L12, L13, L14) that are collected at the end of the second appendix (§6.2) for easy reference.

Below, we start with presenting in §2 the resource-sharing network model, along with the necessary preliminaries, results directly quoted from [20] regarding edges I, II and III in Figure 1, and summarized in Theorem 1. The main result concerning the interchange of limits, edge IV in Figure 1, is presented in §3; refer to Theorem 2. To prove the theorem, the major steps and intermediate results leading to it are detailed in two subsections, §3.1 and §3.2. To facilitate exposition, long proofs of secondary results (organized around Lemma 3 and Proposition 4) are collected in two appendices, §5 and §6.

2. Resource-Sharing Networks The network consists of a set of servers $\mathcal{L}$ and a set of job classes $\mathcal{R}$. Denote $R := |\mathcal{R}|$ and $L := |\mathcal{L}|$, and assume $L \le R$. To be processed in the network, each job of a certain class $r \in \mathcal{R}$ requires the *simultaneous* occupancy of servers, as specified by a *non-negative* matrix of dimension $L \times R$, $A = [a_{\ell r}]_{\ell \in \mathcal{L}, r \in \mathcal{R}}$, Assume A has a full rank (of order L). Denote its ℓ-th row as A_ℓ, a row vector. All other vectors below are column vectors. The superscript, T, of a matrix or vector denotes its transpose.

A special case is when A is an incidence matrix, with $a_{\ell r} = \mathbf{1}\{\ell \in r\}$. Then, A models a (deterministic) routing matrix, with each job class (or "route") r corresponding to a set of "links" (i.e., servers or resources) $\{\ell : a_{\ell r} = 1\}$ that will be concurrently occupied in order to process the jobs in that class. Similarly, the ℓ-th row of the A identifies all the job classes $\{r : a_{\ell r} = 1\}$ that require the service of link ℓ. The general case of A allows randomized (or multi-path) routing.

For each class r, denote the interarrival times between consecutive jobs as $u_r(i)$, and denote the amount of work (service requirement) each job brings to the network as $v_r(i)$, $i = 1, 2, \cdots$. Assume

the interarrival times and work requirements possess finite p-th moments, $p > 2$. In particular, since we need to deal with systems that do not necessarily start empty, we reserve $u_r(1)$ and $v_r(1)$ to denote, at time zero, the *residual* time and work until the next arrival and the next service completion, respectively. Furthermore we assume that $\{(u_r(i), v_r(i)), i \geq 2\}$ are i.i.d. with mean (λ_r^{-1}, ν_r) and variance $(\sigma_{a,r}^2, \sigma_{s,r}^2)$. Denote the offered load (or, traffic intensity) as $\rho = (\rho_r)_{r\in\mathcal{R}}$, with

$$\rho_r = \lambda_r \nu_r. \tag{2}$$

Note that $\lambda_r > 0$ and $\nu_r > 0$ (hence, $\rho_r > 0$) for all $r \in \mathcal{R}$.

The state of the network is $n = (n_r)_{r\in\mathcal{R}}$, where n_r denotes the total number of class r jobs that are present in the network. One job (if any) from each class is processed at any time, while other jobs in the same class waiting in a buffer and will be served on a first-come-first-served basis. Hence, this is a head-of-the-line processor-sharing discipline (with the additional feature of service capacity allocation detailed below).

Each server $\ell \in \mathcal{L}$ has a given capacity, c_ℓ, which is shared among job classes. The allocation of the service capacities takes place in each state, denoted $\Lambda(n) = (\Lambda_r(n))_{r\in\mathcal{R}}$, where $\Lambda_r(n)$ is the capacity allocated to class r when the network state is n. The actual time needed to complete a job then depends on its service requirement and the capacity allocated to it. Specifically, for the i-th class r job mentioned above, provided it is being processed in state n, then the amount of work $v_r(i)$ associated with it is depleted at rate $\Lambda_r(n)$, translating to a service time of $v_r(i)/\Lambda_r(n)$. Let Γ denote the set of all feasible allocations:

$$\Gamma = \{\gamma = (\gamma_r)_{r\in\mathcal{R}} :\ A\gamma \leq c,\ \gamma \geq 0\}. \tag{3}$$

We assume the *proportional fair allocation* is followed, i.e., $\Lambda(n)$ is the solution to the following optimization problem:

$$\max_{\gamma\in\Gamma} \sum_{r\in\mathcal{R}} \beta_r n_r \log(\gamma_r). \tag{4}$$

In the solution, $\Lambda_r(n)$ is unique only for $n_r > 0$. When $n_r = 0$, let $\Lambda_r(n) = 0$, i.e., allocate nothing to class r if the there is no class r present in the network.

The two primitive processes that drive the above network are the *delayed* (i.e., including the residuals) renewal processes associated with the job arrivals and the work or service requirements the jobs bring into the network: $E(t) = (E_r(t))_{r\in\mathcal{R}}$ and $S(t) = (S_r(t))_{r\in\mathcal{R}}$, $t \geq 0$, where

$$E_r(t) = \max\left\{i : \sum_{j=1}^{i} u_r(j) \leq t\right\} \quad \text{and} \quad S_r(t) = \max\left\{i : \sum_{j=1}^{i} v_r(j) \leq t\right\}. \tag{5}$$

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

$$E_r^o(t) = \max\left\{i : \sum_{j=2}^{i} u_r(j) \leq t\right\}, \quad \text{and} \quad S_r^o(t) = \max\left\{i : \sum_{j=2}^{i} v_r(j) \leq t\right\}. \tag{6}$$

Here and below, the superscript "o" denotes the un-delayed version of a (possibly) delayed renewal process.

The two derived processes that characterize, along with the two primitive processes, the dynamics of the network are the queue-length process and the service-completion (departure) process: $N(t) = (N_r(t))_{r\in\mathcal{R}}$ and $D(t) = (D_r(t))_{r\in\mathcal{R}}$, $t \geq 0$, where

$$N_r(t) = N_r(0) + E_r(t) - S_r(D_r(t)), \tag{7}$$
$$D_r(t) = \int_0^t \Lambda_r(N(s))ds. \tag{8}$$

As it will become evident below, it is more convenient to analyze the workload, rather than the queue length, associated with each class:

$$W_r(t) = \nu_r N_r(t), \qquad t \geq 0,\ r \in \mathcal{R}. \tag{9}$$

Similarly, we shall write the generic workload as $w = (w_r)_{r\in\mathcal{R}}$, with the convention $w_r = \nu_r n_r$, throughout below.

We follow the standard approach (e.g., [3, 4, 9, 10]) to construct a Markov process representation of the network by appending to the workload the residual interarrival times and service requirements (at each time instant). Denote $U(t) = (U_r(t))_{r\in\mathcal{R}}$ and $V(t) = (V_r(t))_{r\in\mathcal{R}}$, $t \geq 0$, where:

$$U_r(t) = \sum_{i=1}^{E_r(t)+1} u_r(i) - t, \qquad V_r(t) = \sum_{i=1}^{S_r(D_r(t))+1} v_r(i) - D_r(t). \tag{10}$$

That is, at any given time t, for class r, $U_r(t)$ is the remaining time before the next arrival, and $V_r(t)$ is the remaining service requirement for the job that is in service. (If there is no class r job at the time, $V_r(t)$ is the service requirement for the arriving class r job.) Note, at time $t = 0$, we have $U_r(0) = u_r(1)$ and $V_r(0) = v_r(1)$, the residuals at time zero introduced above. Hence, below we shall refer to $U_r(t)$ and $V_r(t)$ as "residuals" (at t) as well. Then, $\Xi(t) = (W(t), U(t), V(t))$ is a strong Markov process, taking values on the nonnegative orthant of the $3R$-dimensional Euclidean space, denoted $\mathcal{X}$ (cf. [4, 6, 13]). Clearly, the dynamics of the Markov process $\Xi(t)$ 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 $\Xi(t)$ and $W(t)$, wherever necessary we will write $\Xi(t; x)$ and $W(t; x)$, with $x = \Xi(0) \in \mathcal{X}$ being the initial state.

2.1. Diffusion Limit and Preliminary Results To describe the diffusion limit, we introduce a sequence of networks, indexed by k. Each of the networks is like the one introduced above, having the same parameters A, β_r, c_ℓ, and the same allocation $\Lambda(n)$ (hence with the index k omitted from these quantities); but the networks may differ in their arrival rates and mean service times, which will be indexed by k as well. We assume the existence of the following limits of key parameters, as $k \to \infty$:

$$(\lambda_r^k, \nu_r^k, \sigma_{a,r}^k, \sigma_{s,r}^k) \to (\lambda_r, \nu_r, \sigma_{a,r}, \sigma_{s,r}) \ \text{ and } \ k(\rho_r^k - \rho_r) = k(\lambda_r^k \nu_r^k - \lambda_r \nu_r) \to \theta_r, \qquad r \in \mathcal{R}. \tag{11}$$

As a direct consequence of the last convergence, we have $\rho_r^k \to \rho_r$. From now on, we shall specifically regard λ_r, ν_r and ρ_r as the limits defined above, rather than the generic parameters for a particular network as originally introduced.

The limiting regime under diffusion scaling requires a *heavy traffic condition*, which we now specify. A server ℓ is called a bottleneck, if $A_\ell \rho = \sum_{r\in\mathcal{R}} a_{\ell r}\rho_r = c_\ell$, i.e., the total traffic load on that

server is equal to its capacity (asymptotically). Below, for ease of exposition, we shall assume that *all* servers in the network are bottlenecks, and hence, the following heavy traffic condition:

$$A_\ell \rho = c_\ell, \qquad \ell \in \mathcal{L}. \tag{12}$$

Recall, we require the primitives of the network, the interarrival times and service requirements, to possess a finite p-th moment. As now there is a *sequence* of networks, this condition needs to hold *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,

$$\lambda_r^k u_r^k(i) = \lambda_r^1 u_r^1(i) \ \text{ and } \ (\nu_r^k)^{-1} v_r^k(i) = (\nu_r^1)^{-1} v_r^1(i), \quad i \geq 2, \ r \in \mathcal{R}. \tag{13}$$

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

$$\mathsf{E} \sum_{r \in \mathcal{R}} [(u_r^1(2))^p + (v_r^1(2))^p] < \infty. \tag{14}$$

To characterize the diffusion limit below, we follow the same approach in [19] to introduce the fixed-point state space and related matrices (and to the extent possible, use the same notation). Denote $B := \mathrm{diag}(b_r)_{r \in \mathcal{R}}$, with $b_r = \rho_r \nu_r / \beta_r$, which is an R-dimensional diagonal matrix. Associated with the heavy traffic condition is the *fixed-point state space*, denoted as

$$\mathcal{W} = \{w : w = BA^T \pi, \ \pi = (\pi_\ell)_{\ell \in \mathcal{L}} \geq 0\}, \tag{15}$$

which is an L-dimensional polyhedral cone in the positive orthant of the R-dimensional Euclidean space. It is the space where the diffusion limit process (described below) lives. Let $H := (h_m)_{m=1}^{R-L}$ be a matrix that satisfies:

$$ABH = 0 \ \text{ and } \ H^T B H = I. \tag{16}$$

Denote $G := (g_\ell)_{\ell \in \mathcal{L}} := A^T (ABA^T)^{-1}$. Then, we have

$$ABG = I \ \text{ and } \ G^T B H = 0. \tag{17}$$

Any R-dimensional (workload) vector w can be decomposed uniquely as

$$w = BGy + BHz. \tag{18}$$

For any state w, we can measure its distance from the fixed-point state space $\mathcal{W}$ as follows:

$$d^{fp}(w) = \sum_{\ell \in \mathcal{L}} (-g_\ell^T w)^+ + \sum_{m=1}^{R-L} |h_m^T w|. \tag{19}$$

Observe that for any R-dimensional vector w, $w \in \mathcal{W}$ if and only if $G^T w = \pi \geq 0$ and $H^T w = 0$. Therefore, w is a fixed-point state (or an invariant state) if and only if $d^{fp}(w) = 0$,

The Markov process associated with the k-th network is $\Xi^k(t) = (W^k(t), U^k(t), V^k(t))$, and it follows the dynamics in (7-10) with the index k suitably appended.

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

$$(\hat{E}_r^{o,k}(t), \hat{S}_r^{o,k}(t)) = \frac{1}{k}\left(E_r^{o,k}(k^2t) - \lambda_r^k k^2 t, S_r^{o,k}(k^2t) - (\nu_r^k)^{-1}k^2t\right), \tag{20}$$
$$(\hat{E}_r^{k}(t), \hat{S}_r^{k}(t)) = \frac{1}{k}\left(E_r^{k}(k^2t) - \lambda_r^k k^2 t, S_r^{k}(k^2t) - (\nu_r^k)^{-1}k^2t\right),$$
$$(\hat{\Xi}_r^{k}(t), \hat{N}_r^{k}(t), \hat{W}_r^{k}(t)) = \frac{1}{k}\left(\Xi_r^{k}(k^2t), N_r^{k}(k^2t), W_r^{k}(k^2t)\right).$$

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

$$(\bar{E}_r^{o,k}(t), \bar{S}_r^{o,k}(t), \bar{E}_r^{k}(t), \bar{S}_r^{k}(t)) = \frac{1}{k}\left(E_r^{o,k}(kt), S_r^{o,k}(kt), E_r^{k}(kt), S_r^{k}(kt)\right). \tag{21}$$

Under the diffusion scaling, we can rewrite the dynamics in (7-9) as follows:

$$\begin{aligned}\hat{W}^k(t) &= \hat{W}^k(0) + \hat{X}^k(t) + k[\rho t - \tilde{D}^k(t)] \\ &= \hat{W}^k(0) + \hat{X}^k(t) + BG\hat{Y}^k(t) + BH\hat{Z}^k(t); \end{aligned}\tag{22}$$
$$\tilde{D}^k(t) = \int_0^t \Lambda(\hat{N}^k(s))ds; \tag{23}$$
$$\hat{X}_r^k(t) = \nu_r^k\left(\hat{E}_r^k(t) - \hat{S}_r^k(\tilde{D}_r^k(t))\right) + k(\rho_r^k - \rho_r)t, \quad \text{for } r \in \mathcal{R}; \tag{24}$$
$$\hat{Y}^k(t) = kA[\rho t - \tilde{D}^k(t)] = k[ct - A\tilde{D}^k(t)], \text{ is non-decreasing in } t \geq 0, \quad \text{for } \ell \in \mathcal{L}; \tag{25}$$
$$\hat{Z}^k(t) = kH^T[\rho t - \tilde{D}^k(t)]. \tag{26}$$

The process, $\tilde{D}_r^k(t) = D_r^k(k^2t)/k^2$, is a variation of what's known as the fluid-scaled process $\bar{D}_r^k(t)$ $(= D_r^k(kt)/k)$. The balance equation (the second equality in (22)) follows from the decomposition in (18). Denote $\hat{X}^k(t) = (\hat{X}_r^k(t))_{r\in\mathcal{R}}$.

For the derived processes, denote their limits as follows:

$$\hat{W}(t) = (\hat{W}_r(t))_{r\in\mathcal{R}}, \quad \hat{X}(t) = (\hat{X}_r(t))_{r\in\mathcal{R}}, \quad \hat{Y}(t) = (\hat{Y}_\ell(t))_{\ell\in\mathcal{L}}, \quad \hat{Z}(t) = (\hat{Z}_m(t))_{m=1}^{R-L}.$$

Furthermore, the limiting processes are characterized by the following so-called *dynamic complementarity problem* (DCP):

$$\hat{W}(t) = \hat{W}(0) + \hat{X}(t) + BG\hat{Y}(t) + BH\hat{Z}(t) \ (\geq 0), \quad \text{for } t \geq 0; \tag{27}$$
$$G^T\hat{W}(t) \geq 0, \quad \text{for } t \geq 0; \tag{28}$$
$$\hat{Y}_\ell(t) \text{ is non-decreasing in } t \geq 0, \ \hat{Y}_\ell(0) = 0, \quad \ell \in \mathcal{L}; \tag{29}$$
$$\int_0^\infty \hat{W}(t)^T G \ d\hat{Y}(t) = 0; \tag{30}$$
$$H^T\hat{W}(t) = 0, \quad \text{for } t \geq 0; \tag{31}$$
$$\hat{Z}(0) = 0; \tag{32}$$

where $\hat{W}(0)$ is the (given) initial state and $\hat{X}(t)$, the "free process," is a Brownian motion with drift (vector) $\theta = (\theta_r)_{r\in\mathcal{R}}$ specified in (11), and covariance (matrix)

$$\Upsilon = \text{diag}(\sigma_r^2)_{r\in\mathcal{R}}, \qquad \text{with } \ \sigma_r^2 = \nu_r^2(\lambda_r^3\sigma_{a,r}^2 + \rho_r\nu_r^{-3}\sigma_{s,r}^2) = \lambda_r\nu_r^2(\lambda_r^2\sigma_{a,r}^2 + \nu_r^{-2}\sigma_{s,r}^2). \tag{33}$$

The existence of the limiting processes defined above is part of the next theorem. Indeed, in [19], we have established that for any given free process $\hat{X}(t)$ that is right continuous with left limits (RCLL), there exists a unique (pathwise) solution, $(\hat{W}(t), \hat{Y}(t), \hat{Z}(t))$, to the DCP in (27-32). Moreover, we will study the weak convergence (denoted as "$\Rightarrow$") 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, 18]). 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.

The following DCP is a deterministic version of the one in (27-32), with the free process (unreflected Brownian motion) $\hat{X}(t)$ replaced by its drift term θt:

$$\hat{w}(t) = \hat{w}(0) + \theta t + BG\hat{y}(t) + BH\hat{z}(t) \ (\geq 0), \quad \text{for } t \geq 0; \tag{34}$$
$$G^T\hat{w}(t) \geq 0, \quad \text{for } t \geq 0; \tag{35}$$
$$\hat{y}_\ell(t) \text{ is non-decreasing in } t \geq 0, \ \hat{y}_\ell(0) = 0, \quad \ell \in \mathcal{L}; \tag{36}$$
$$\int_0^\infty \hat{w}(t)^T G \ d\hat{y}(t) = 0; \tag{37}$$
$$H^T\hat{w}(t) = 0, \quad \text{for } t \geq 0; \tag{38}$$
$$\hat{z}(0) = 0. \tag{39}$$

Below, we shall refer to the deterministic DCP in (34-39) as *stable*, if there exists a time T such that for any solution with $|\hat{w}(0)| \leq 1$, we have $\hat{w}(t) = 0$ for all $t \geq T$. It is known (Theorem 8(a) of [20]) that the deterministic DCP is stable if and only if

$$A\theta < 0.$$

Theorem 1 (a) (Edge I [19]) Suppose the heavy traffic condition in (12) is in force; and under the diffusion scaling, the initial workloads converge to some (random) fixed-point state, while the (time-zero) residuals vanish:

$$\hat{W}^k(0) \Rightarrow \hat{W}(0) \in \mathcal{W}, \tag{40}$$
$$|\hat{U}^k(0)| + |\hat{V}^k(0)| = \frac{1}{k}(|u^k(1)| + |v^k(1)|) \to 0. \tag{41}$$

Then, the following weak convergence holds when $k \to \infty$:

$$\left(\hat{W}^k(\cdot), \hat{X}^k(\cdot), \hat{Y}^k(\cdot), \hat{Z}^k(\cdot)\right) \Rightarrow \left(\hat{W}(\cdot), \hat{X}(\cdot), \hat{Y}(\cdot), \hat{Z}(\cdot)\right),$$

with the limit characterized by the equations in (27-32).
(b) (Edge III [7, 20]) If the deterministic DCP in (34-39) is stable (or equivalently, $A\theta < 0$), then the diffusion limit $\hat{W}(t)$ in part (a) above is positive recurrent and has a unique stationary distribution.
(c) (Edge II [5, 20]) Suppose the heavy traffic condition in (12) holds, and the DCP in (34-39) is stable. Then, for any sufficiently large k, $\hat{\Xi}^k(t) = (\hat{W}^k(t), \hat{U}^k(t), \hat{V}^k(t))$ is positive (Harris) recurrent and has a unique stationary distribution. Furthermore, if the p-th moment condition in (14) holds, then for any $m \in [0, p-1)$ and for sufficiently large k, the stationary workload has a finite m-th moment and

$$\lim_{t\to\infty} \mathsf{E}|\hat{W}^k(t;x)|^m = \mathsf{E}|\hat{W}^k(\infty)|^m < \infty, \quad \text{for any initial state } x, \tag{42}$$

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

3. Interchange of Limits We are now ready to establish edge IV in Figure 1, along with the convergence of the moments of the stationary workloads, We need the following condition.

p^*-th moment condition: All interarrival times and service requirements have bounded p^*-th moments, i.e., for some $p^* > 2(p+2)$,

$$\mathsf{E}\sum_{r\in\mathcal{R}}[(u_r^1(2))^{p^*} + (v_r^1(2))^{p^*}] < \infty. \tag{43}$$

As will be shown below, this condition will guarantee the boundedness of the p-th moment of the workload, which holds the key to proving the convergence of stationary m-th moments of the 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 $t \geq 0$,

$$\mathsf{E}\sup_{0\leq s\leq t}\sum_{r\in\mathcal{R}}\left(|E_r^{o,k}(s) - \lambda_r^k s|^{p^*} + |S_r^{o,k}(s) - \mu_r^k s|^{p^*}\right) \leq \kappa(1 + t^{p^*/2}), \tag{44}$$

and furthermore,

$$\mathsf{E}\sup_{0\leq s\leq t}\sum_{r\in\mathcal{R}}\left(|\hat{E}_r^{o,k}(s)|^{p^*} + |\hat{S}_r^{o,k}(s)|^{p^*}\right) \leq \kappa(1 + t^{p^*/2}). \tag{45}$$

The above variation is technically convenient and has been used in previous studies [3, 20]. To prove the claimed implication, refer to Appendix 5.1, Lemma 9.

Theorem 2 Assume the heavy traffic condition in (12), the stability of the deterministic DCP in (34-39) (i.e., $A\theta < 0$), and the p^*-th moment condition in (43). Then, the following weak convergence of stationary distributions holds,

$$\hat{W}^k(\infty) \Rightarrow \hat{W}(\infty), \quad \text{as } k \to \infty. \tag{46}$$

In particular, since $\hat{W}(\infty)$ follows the stationary distribution of $\hat{W}(t)$ as $t \to \infty$, the interchange of the limits, $t \to \infty$ and $k \to \infty$, illustrated in Figure 1 (edges III and IV) is valid. Furthermore, for any $m \in [0, p-1)$,

$$\mathsf{E}|\hat{W}^k(\infty)|^m \to \mathsf{E}|\hat{W}(\infty)|^m, \quad \text{as } k \to \infty. \tag{47}$$

The above theorem parallels Theorem 14 in [20], which also justifies the interchange of limits but under a different condition, the so-called (pathwise) bounded workload condition: for some constant $\kappa > 0$,

$$\sup_{0\leq s\leq t}|\hat{W}^k(s)| \leq \kappa\left(|\hat{W}^k(0)| + \sup_{0\leq s\leq t}|\hat{X}^k(s)|\right). \tag{48}$$

In particular, Theorem 14 of [20] shows the above implies the boundedness of the p-th moment of the workload process,

$$\mathsf{E}\sup_{0\leq s\leq t}|\hat{W}^k(s)|^p \leq \kappa(|\hat{\Xi}^k(0)|^p + 1 + t^p), \tag{49}$$

under the p-th moment condition of primitives in (14). Clearly, the above bounded workload condition and the p^*-th moment condition do not imply each other: the latter is trivial to verify, being imposed directly on the primitives; whereas the former does not require higher (than order p) moments on the primitives but could be difficult (at least non-trivial) to verify.

To prove Theorem 2, we break into two subsections below. First, in §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. We develop bounds on the probabilities of these regular events, and demonstrate that the workload process, too, behaves "nicely" under these events. For the latter, we apply Bramson's "hydro-dynamic" approach ([2, 19]) to show that the bounds in (48) work for sample paths under the regular events. Next, combining these results in §3.2 with a (crude) bound for the p-th moment of the workload under the "non-regular" (or, rare) events, we derive the p-th moment bound of the workload in (49) under the diffusion scaling (Lemma 5). The rest is then similar to the steps in [20], i.e., establishing the key properties of the workload process (such as uniform integrability, uniform p-th moment stability and tightness), which then complete the proof of Theorem 2.

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

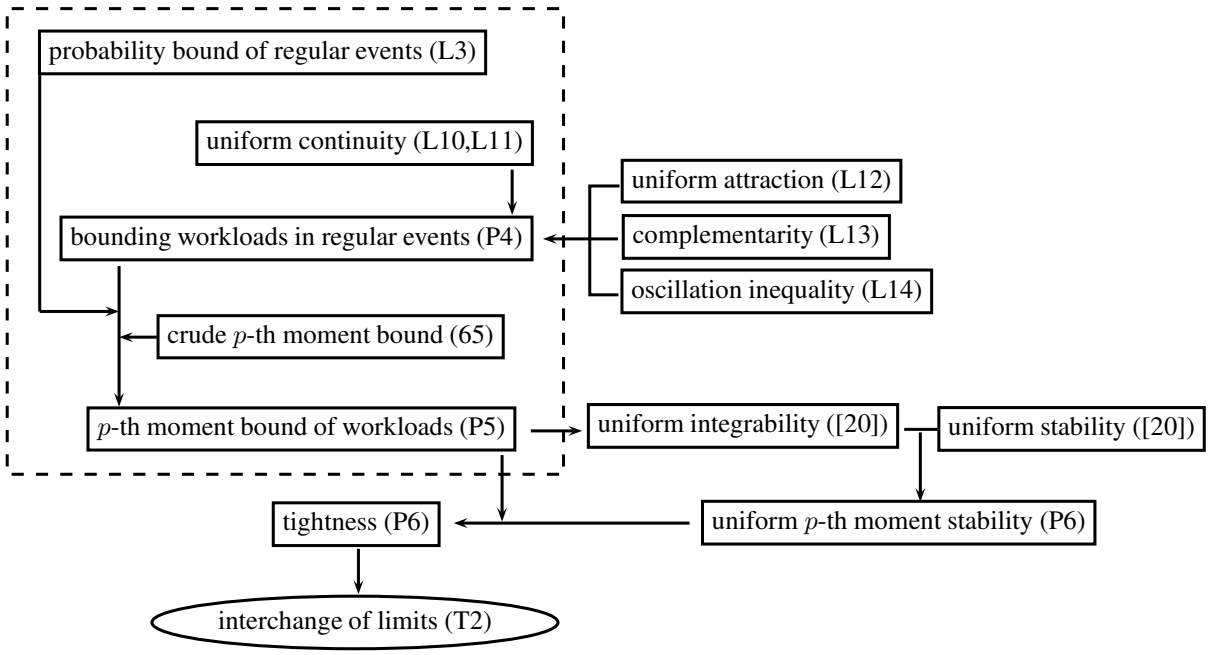

FIGURE 2. Roadmap (T, P, L: Theorem, Proposition, Lemma)

3.1. Workload under Regular Events Define the variables:

$$u_r^{k,\max}(t) := \max\left\{ u_r^k(i) : \sum_{i'=2}^{i-1} u_r^k(i') \le t,\ i = 2,3,\cdots \right\}, \tag{50}$$

$$v_r^{k,\max}(t) := \max\left\{ v_r^k(i) : \sum_{i'=2}^{i-1} v_r^k(i') \le t,\ i = 2,3,\cdots \right\}. \tag{51}$$

The first variable is the maximal interarrival time of class r realized before time t for the k-th 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 times, and $\{m_k\}_{k\in\mathcal{K}}$ be a sequence of real numbers with $m_k \geq 1$. Define the regular events as

$$\Omega^k(t^*,u^*,m_k) = \Omega_u^k(t^*,m_k)\cap\Omega_v^k(t^*,m_k)\cap\Omega_X^k(t^*,m_k)\cap\Omega_E^k(t^*,u^*,m_k)\cap\Omega_S^k(t^*,u^*,m_k), \quad (52)$$

where

$$\Omega_u^k(t^*,m_k) = \bigcap_{r\in\mathcal{R}}\left\{\frac{1}{km_k}u_r^{k,\max}(k^2m_kt^*)\leq\frac{1}{k^{(p^*-2)/2p^*}}\right\}, \quad (53)$$

$$\Omega_v^k(t^*,m_k) = \bigcap_{r\in\mathcal{R}}\left\{\frac{1}{km_k}v_r^{k,\max}(k^2m_kt^*)\leq\frac{1}{k^{(p^*-2)/2p^*}}\right\}, \quad (54)$$

$$\Omega_E^k(t^*,u^*,m_k) = \left\{\sup_{0\leq t\leq kt^*}\sup_{0\leq u\leq u^*}|\frac{1}{m_k}(\bar{E}^{o,k}(m_k(t+u))-\bar{E}^{o,k}(m_kt))-\lambda^ku|\leq\frac{1}{\log k}\right\}, \quad (55)$$

$$\Omega_S^k(t^*,u^*,m_k) = \left\{\sup_{0\leq t\leq kt^*}\sup_{0\leq u\leq u^*}\frac{1}{m_k}|(\bar{S}^{o,k}(m_k(t+u))-\bar{S}^{o,k}(m_kt))-\mu^ku|\leq\frac{1}{\log k}\right\}, \quad (56)$$

$$\Omega_X^k(t^*,m_k) = \left\{\sup_{0\leq t\leq t^*}\frac{1}{m_k}(|\hat{E}^{o,k}(m_kt)|+|\hat{S}^{o,k}(m_kt)|)\leq\frac{k}{\log k}\right\}. \quad (57)$$

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 §5.

Lemma 3 Let t^* and u^* be any positive times. Then, the following estimate holds for sufficiently large k (depending on t^* and u^*),

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

Next, denote

$$y^k[=y^k(\omega,\Delta,m_k)] := \max\left(\frac{1}{m_k}|\hat{W}^k(0)|+\sup_{0\leq t\leq\Delta}\frac{1}{m_k}|\hat{X}^k(m_kt)|,\ \frac{1}{m_k}|\hat{\Xi}^k(0)|,\ 1\right), \quad (58)$$

for any time interval $[0,\Delta]$, with $\Delta>0$, and any sequence of numbers $\{m_k\geq 1; k\in\mathcal{K}\}$. 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 $\lceil k\Delta/y^kT\rceil$ segments with equal length y^kT/k, where $\lceil\cdot\rceil$ denotes the integer ceiling. Observe that for any $\omega\in\Omega_X^k$, the amount of segments $k\Delta/y^kT\geq O(\log k)\to\infty$ as $k\to\infty$. The j-th segment, $j=0,...,\lceil k\Delta/y^kT\rceil-1$, covers the time interval $[jy^km_kT/k,(j+1)y^km_kT/k]$. Note that the last interval (with $j=\lceil k\Delta/y^kT\rceil-1$) covers a negligible piece of time beyond the right end of $[0,\Delta]$ if $k\Delta/y^kT$ is not an integer. For simplicity, below we shall treat $k\Delta/y^kT$ as an integer so as to omit the ceiling notation. Then, for any $t\in[0,\Delta]$, we can write $t=y^k(jT+u)/k$ for some $j=0,\cdots,k\Delta/y^kT$ and $u\in[0,T]$. Therefore, for $u\in[0,T]$ and $j\leq k\Delta/y^kT$, we write

$$\begin{aligned}\frac{1}{y^km_k}\hat{W}^k(m_kt) &= \frac{1}{y^km_k}\hat{W}^k(\frac{jy^km_kT+y^km_ku}{k})\\ &= \frac{1}{ky^km_k}W^k(jky^km_kT+ky^km_ku) := \bar{W}^{k,j}(u).\end{aligned} \quad (59)$$

The definition for the "hydro-dynamic" scaling above, $\bar{W}^{k,j}(u)$, is slightly different from the same notation in our previous papers (e.g., [19]) in that new parameters m_k and y^k are introduced into the scaling. Also note that we have suppressed the parameters m_k and y^k in the above hydro-dynamic scaling for ease of notation. Some other processes, $\bar{\Xi}^{k,j}(u)$, $\bar{Y}^{k,j}(u)$, $\bar{U}^{k,j}(u)$ and $\bar{V}^{k,j}(u)$, are defined in the same manner. For convenience, we define the maximum and minimum operators as $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$, respectively; and when a and b are vectors, the operators apply component-wise.

Proposition 4 Consider any time interval $[0, \Delta]$, with $\Delta > 0$, and let $\epsilon > 0$ be any given (small) number. Pick a sufficiently large T, and define (for convenience) the following constants:

$$\bar{\Delta} = (\max\{c_\ell\} \vee 1)\Delta + 1, \quad \bar{T} = (\max\{c_\ell\} \vee 1)T. \tag{60}$$

Then, for sufficiently large k, the following properties hold for any initial state $\hat{\Xi}^k(0)$, $m_k \geq (|\hat{\Xi}^k(0)| \vee 1)$, $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, and *positive* integers $j = 1, \cdots, k\Delta/y^kT$:

(a) (uniform attraction)

$$d^{fp}(\bar{W}^{k,j}(u)) \leq \epsilon, \quad \text{for all } u \in [0, T]; \tag{61}$$

(b) (complementarity) if $g_\ell^T \bar{W}^{k,j}(u') > \epsilon$ for some $u' \in [0, T]$, then

$$\bar{Y}_\ell^{k,j}(u) - \bar{Y}_\ell^{k,j}(0) = 0, \quad \text{for all } u \in [0, T];$$

(c) (boundedness)

$$|\bar{W}^{k,j}(u)| \leq \kappa, \quad \text{for all } u \in [0, T], \tag{62}$$

where κ is a positive constant that depends only on network parameters, i.e., independent of k and ω. In addition, the boundedness in (62) also applies to $j = 0$.

The proof of the proposition is deferred to the appendix, §6. Note this proposition strengthens Lemma 7 of [19] in that the results in (a-c) here hold uniformly on the regular events and allow for additional scaling parameters. In fact, the result in (c) is what we need to bound the p-th moment of the workload below, while those in (a) and (b) are auxiliary properties. Specifically, from the definitions in (58, 59), the inequality in (c) reads $|\hat{W}^k(m_k t)/m_k| \leq \kappa y^k$ for $t \in [0, \Delta]$, that is, the workload is dominated by the free process plus the initial workload for sample paths in the regular events. This bound serves as an input to the next proposition in deriving the p-th moment of the workload.

3.2. Key Properties Leading to the Interchange Bounding the p-th moment of the workload is a crucial step in justifying the interchange of limits. To do so, we cannot simply follow standard approaches in the literature, e.g., Dai ([4], Lemma 4.5) and Dai and Meyn ([5], 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. Under diffusion scaling (which is required in our setting), the arrival processes become unbounded as $k \to \infty$ and cannot serve the same purpose. To overcome this difficulty, our approach is to identify the regular events and characterize the hydro-dynamics of the networks under these events as summarized in Lemma 3 and Proposition 4 above. Equipped with these results, we are ready to bound the p-th moment of the workload, as stated in the proposition below.

Proposition 5 (Bounded p-th Moment of Workload) There is a constant $\kappa > 0$ such that for any time $t \geq 0$ and sufficiently large k, the following holds for any initial state $\hat{\Xi}^k(0)$ and any $m_k \geq (|\hat{\Xi}^k(0)| \vee 1)$

$$\mathsf{E} \sup_{0\leq s\leq t} \left| \frac{1}{m_k} \hat{W}^k(m_k s) \right|^p \leq \kappa(1+t^p). \tag{63}$$

Proof. Following Proposition 4, we first bound the workload processes in $\Omega^k(\bar{\Delta}, \bar{T}, m_k)$. By Proposition 4(c), 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 $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$:

$$\sup_{0\leq t\leq \Delta} \left| \frac{1}{m_k} \hat{W}^k(m_k t) \right| \leq \kappa_1 y^k \leq \kappa_1 \left(\frac{|\hat{W}^k(0)|}{m_k} + \sup_{0\leq s\leq \Delta} \frac{|\hat{X}^k(m_k s)|}{m_k} + \frac{|\hat{\Xi}^k(0)|}{m_k} + 1 \right). \tag{64}$$

Hence, we have

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

where the last inequality is proved following the same procedure as in Lemma 9(a) of [20]. Next, we bound the workload in $\Omega \setminus \Omega^k(\bar{\Delta}, \bar{T}, m_k)$. 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,

$$\begin{aligned}
&\mathsf{E} \left[\left(\sup_{0\leq t\leq \Delta} \left| \frac{1}{m_k} \hat{W}^k(m_k t) \right| \right)^p \cdot 1_{\Omega\setminus\Omega^k(\bar{\Delta},\bar{T},m_k)} \right] \\
&\leq \left[\mathsf{E} \left(\sup_{0\leq t\leq \Delta} \left| \frac{1}{m_k} \hat{W}^k(m_k t) \right| \right)^{\alpha p} \right]^{\frac{1}{\alpha}} \cdot \left[\mathsf{E} \left(1_{\Omega\setminus\Omega^k(\bar{\Delta},\bar{T},m_k)} \right)^{\beta} \right]^{\frac{1}{\beta}} \\
&\leq \left[\mathsf{E}\kappa_2 \left(1 + \left| \frac{1}{km_k} E^{o,k}(k^2 m_k \Delta) \right|^{\alpha p} \right) \right]^{\frac{1}{\alpha}} \cdot \left[\mathsf{P}(\Omega \setminus \Omega^k(\bar{\Delta}, \bar{T}, m_k)) \right]^{\frac{1}{\beta}} \\
&\leq \left[\kappa_3 \left(1 + (k\Delta)^{\alpha p} \right) \right]^{\frac{1}{\alpha}} \cdot \left[\frac{(\log k)^{p^*+1}}{k^{p^*/2-2}} \right]^{\frac{1}{\beta}} \\
&\leq \kappa_4(1+\Delta^p).
\end{aligned} \tag{65}$$

We have applied Holder's inequality and Lemma 3 in the first and fourth 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,

$$\begin{aligned}
\frac{1}{m_k} \hat{W}_r^k(m_k t) &= \frac{1}{m_k} \hat{W}_r^k(0) + \frac{1}{km_k} \nu_r^k E_r^k(k^2 m_k t) - \frac{1}{km_k} \nu_r^k S_r^k(D_r^k(k^2 m_k t)) \\
&\leq 1 + \frac{1}{km_k} \nu_r^k [1 + E_r^{o,k}(k^2 m_k t)] \leq 1 + \nu_r^k + \frac{1}{km_k} \nu_r^k E_r^{o,k}(k^2 m_k t).
\end{aligned}$$

Finally, the above two estimates lead to the proposition, since the time Δ is arbitrarily given in Proposition 4. □

Compared against Lemma 9 of [20], here we have removed the so-called bounded workload condition, but at the expense of requiring the higher, p^*-th moment condition on the primitives, in establishing the p-th moment bound on the workload.

Given the p-th moment bound of the workload in the above proposition, the remaining building blocks to establish the interchange of limits are rather standard, as illustrated in Figure 2, the

blocks following "p-th moment bound of workloads." First, this will imply the uniform integrability of the workload processes, which, along with a uniform stability property previously established (Theorem 7(a) of [20]), will lead to the uniform p-th moment stability and tightness. With the tightness property the proof of our main result here, Theorem 2, then follows standard arguments typically used in justifying the interchange of limits ([3, 9, 10, 13, 14]). For completeness, we sketch these steps below.

First, note that Proposition 5 holds with p replaced by any p' satisfying $p < p' < p^*/2 - 2$, which implies the *uniform integrability* of workload (Lemma 9(c) of [20]), i.e., for any $t > 0$,

$$\{|\hat{W}^k(m_k t)/m_k|^p\}_k \text{ is uniformly integrable.} \tag{66}$$

Observe from the diffusion limit (Theorem 1(a)) that $\hat{W}^k(m_k t)/m_k$ should be close to $\hat{W}(m_k t)/m_k$ and should approximate $\hat{w}(t)$ (with $\hat{w}(t) \le 1$), the deterministic counterpart of $\hat{W}(t)$ as defined in (34-39). If $\hat{w}(t)$ is stable (i.e., there exists a constant time t_0 such that $\hat{w}(t) = 0$ for $t \ge t_0$), then for any $t \ge t_0$, $\hat{W}^k(m_k t)/m_k$ should be close to 0 for sufficiently large k. Indeed, we can establish

$$\frac{1}{m_k}\hat{W}^k(m_k t) \to 0 \quad \text{ u.o.c. of } t \ge t_0,$$

where $m_k \ge |\hat{\Xi}^k(0)|$ and $m_k \to \infty$ as $k \to \infty$. Consequently, for any $t \ge t_0$, we have

$$\lim_{k\to\infty} \mathsf{E}\frac{1}{m_k^p}\left|\hat{W}^k(m_k t)\right|^p = \mathsf{E}\lim_{k\to\infty}\frac{1}{m_k^p}\left|\hat{W}^k(m_k t)\right|^p = 0,$$

where the interchange of the expectation and the limit in the first equality is justified by the uniform integrability in (66). Furthermore, applying the *uniform stability* property (Theorem 7(a) of [20]), we can strengthen the above to obtain the *uniform p-th moment stability*. These results are summarized in part (a) of the proposition below. Part (b) of the proposition, the tightness property, then follows from part (a) and standard approaches as in [3, 5, 20]. (Thus, no proof is required.)

Proposition 6 Assume the heavy traffic condition in (12), the stability of the deterministic DCP in (34-39), and the p^*-th moment condition in (43).
(a) (Uniform p-th Moment Stability) There exists t_0 such that the following holds for all $t \ge t_0$,

$$\lim_{|x|\to\infty} \sup_k \mathsf{E}\frac{1}{|x|^p}\left|\hat{W}^k(|x|t; x)\right|^p = 0. \tag{67}$$

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

Finally, given the tightness property in the above proposition, the proof of our main result here, Theorem 2, is identical to the proof of Theorem 4 in [20], which we outline below for completeness.

In a nutshell, the argument starts by initializing the process $\hat{\Xi}^k(t)$ $(=(\hat{W}^k(t), \hat{U}^k(t), \hat{V}^k(t)))$ in its stationary distribution $\hat{\pi}^k$ $(=(\hat{\pi}^k_1, \hat{\pi}^k_2, \hat{\pi}^k_3))$. It is unclear whether this initialization satisfies the condition in (40); hence, Theorem 1 cannot be applied as yet. Nevertheless, since $\{\hat{\pi}^k\}$ is tight, we can establish a variation of Theorem 1 (refer to Proposition 2 of [20]), such that for a subsequence $\mathcal{K}$, $\{\hat{W}^k(t^k_0 + t); k \in \mathcal{K}\}$ (with t^k_0 being a carefully chosen sequence of times that approaches 0 as $k \to \infty$) converges weakly to a limit $\hat{W}(t)$, as characterized in (27-32). The choice of initialization makes $\hat{W}^k(t^k_0 + t)$ equal in distribution to $\hat{W}^k(0)$, for each $k \in \mathcal{K}$. Hence, as $k \to \infty$, the limit $\hat{W}(t)$ follows the same distribution as that of $\hat{W}(0)$, namely, $\tilde{\pi}_1$. Consequently, $\hat{W}(t)$ follows the distribution $\tilde{\pi}_1$

for all $t \geq 0$, which implies that $\tilde{\pi}_1$ must coincide with the *unique* stationary distribution $\hat{\pi}_1$ of the limit $\hat{W}(t)$ guaranteed by Theorem 1(c).

In summary, we can conclude that $\hat{\pi}_1$ is the weak limit of any convergent subsequence of $\{\hat{\pi}_1^k\}$. Thus, the full sequence $\{\hat{\pi}_1^k\}$ must converge weakly to $\hat{\pi}_1$, which is the weak convergence in (46). Finally, the convergence in (47) is a direct consequence of the convergence in (46) and the "furthermore" part in Proposition 6(b).

4. Concluding Remarks In a prior study [20], we have provided a justification of the diffusion approximation, via the proof of the interchange of limits, under the bounded workload condition. Here, we have replaced this condition by a direct moment condition on the primitives, but at the expense of requiring the higher, p^*-th moment condition on the interarrival and service times, in establishing the p-th moment bound on the workload, with $p^* > 2(p+2)$. Together the two studies provide a thorough investigation on the interchange of limits in resource sharing networks, under two different but complementary sets of conditions. In addition, they have also revealed some of the key differences between resource sharing networks and the more traditional multiclass queueing networks.

5. Appendix: Proof of Lemma 3 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 §5.1 after the proof.

To lighten notation, below we shall omit m_k from the scaling parameters (not the indices) km_k and $k^2 m_k$. This can be equivalently viewed as setting $m_k = 1$ for all k, which is innocuous, since $m_k \geq 1$. In the same spirit, we also omit the arguments t^*, u^* and m_k associated with the relevant events.

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 [2]. 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 $\{u_r^{k,\max}(k^2 m_k t^*)/km_k \leq 1/k^{(p^*-2)/2p^*}\}$ as the number of job classes is finite. Consider any fixed r, and pick a sufficiently large constant $\kappa_1 > 0$. We have

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

The first term F_1 is estimated by applying Lemma 8 and the Markov inequality as follows:

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

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

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

The above estimates yield that there is a constant κ_3, independent of k, such that for sufficiently large k (and for any $m_k \ge 1$),

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

Summing up the above over all r, we have for sufficiently large k (and for any $m_k \ge 1$),

$$\mathsf{P}(\Omega \setminus \Omega_u^k) \le \frac{\kappa_4}{k^{p^*/2-1}}. \tag{68}$$

Probability bound for Ω_X^k Note that there exists a constant κ_1 such that

$$\left(\sup_{0\le t\le t^*} \frac{1}{m_k} (|\hat{E}^{o,k}(m_k t)| + |\hat{S}^{o,k}(m_k t)|) \right)^{p^*} \le \kappa_1 \sup_{0\le s\le t^*} \sum_{r\in\mathcal{R}} \left(|\frac{\hat{E}_r^{o,k}(m_k s)}{m_k}|^{p^*} + |\frac{\hat{S}_r^{o,k}(m_k s)}{m_k}|^{p^*} \right).$$

Applying the above inequality, the p^*-th moment condition in (45), Markov inequality, and Lemma 9, we have the following estimation (for all $m_k \ge 1$),

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

Probability bounds for Ω_E^k and Ω_S^k We estimate the probability bound for Ω_E^k only, since the bound for Ω_S^k is 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^*),

$$\begin{aligned} J &:= \mathsf{P}\left(\left\{ |\frac{1}{m_k} (\bar{E}_r^{o,k}(m_k(t+u)) - \bar{E}_r^{o,k}(m_k t)) - \lambda_r^k u| > \frac{1}{2\alpha \log k} \right\} \bigcap \Omega_u^k \right) \\ &\le \kappa_1 \frac{(\alpha \log k)^{p^*}}{k^{p^*/2-1}}. \end{aligned} \tag{69}$$

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 (55). Write the term involving the arrival process in (69) as,

$$|\frac{1}{m_k} (\bar{E}_r^{o,k}(m_k(t+u)) - \bar{E}_r^{o,k}(m_k t)) - \lambda_r^k u|$$

$$
\begin{aligned}
&\le |\frac{1}{m_k}(\bar{E}_r^{o,k}(m_kt+\bar{U}_r^k(m_kt)+m_ku)-\bar{E}_r^{o,k}(m_kt+\bar{U}_r^k(m_kt)))-\lambda_r^ku|\\
&\quad+\frac{1}{m_k}|\bar{E}_r^{o,k}(m_kt+\bar{U}_r^k(m_kt)+m_ku)-\bar{E}_r^{o,k}(m_kt+m_ku)|\\
&\quad+\frac{1}{m_k}|\bar{E}_r^{o,k}(m_kt+\bar{U}_r^k(m_kt))-\bar{E}_r^{o,k}(m_kt)|.
\end{aligned}
$$

Observe in the first term at the right-hand-side of the above that $J_1 := (\bar{E}_r^{o,k}(m_kt+\bar{U}_r^k(m_kt)+m_ku)-\bar{E}_r^{o,k}(m_kt+\bar{U}_r^k(m_kt)))/m_k-\lambda_r^ku$ and $(\bar{E}_r^{o,k}(m_ku)/m_k-\lambda_r^ku)$ are equal in distribution. This is because that the time $m_kt+\bar{U}_r^k(m_kt)$ 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 $\bar{U}_r^k(m_kt)$, the third term is equal to $1/(km_k)$ $(\le 1/k)$. For the middle term, we restrict our attention to $\omega\in\Omega_u^k$, which implies $\bar{U}_r^k(m_kt)\le 1/k^{(p^*-2)/2p^*}$. Then, we have the following estimate on this term,

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

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

$$
\left|\frac{1}{m_k}(\bar{E}_r^{o,k}(m_k(t+u))-\bar{E}_r^{o,k}(m_kt))-\lambda_r^ku\right|\le J_1+J_2+\frac{2}{k},\quad t\in[0,kt^*],\ u\in[0,u_1^*],
$$

and consequently,

$$
\begin{aligned}
J &\le \mathsf{P}\left(\left\{|J_1|+|J_2|+\frac{2}{k}>\frac{1}{2\alpha\log k}\right\}\bigcap\Omega_u^k\right)\\
&\le \mathsf{P}\left\{|\frac{1}{m_k}\bar{E}_r^{o,k}(m_ku)-\lambda_r^ku|>\frac{1}{4\alpha\log k}-\frac{1}{k}\right\}\\
&\quad+\mathsf{P}\left\{\frac{1}{m_k}\bar{E}_r^{o,k}(1/k^{(p^*-2)/2p^*})>\frac{1}{4\alpha\log k}-\frac{1}{k}\right\}\\
&\le \frac{\mathsf{E}|\bar{E}_r^{o,k}(m_ku)/m_k-\lambda_r^ku|^{p^*}}{(1/4\alpha\log k-1/k)^{p^*}}+\frac{\mathsf{E}[\bar{E}_r^{o,k}(1/k^{(p^*-2)/2p^*})/m_k]^{p^*}}{(1/4\alpha\log k-1/k)^{p^*}}\\
&\le \frac{\kappa'(1/k^{p^*})(1+(ku_1^*)^{p^*/2})}{(1/4\alpha\log k-1/k)^{p^*}}+\frac{\kappa''(1/k^{p^*})\left(k/k^{(p^*-2)/2p^*}\right)^{p^*}}{(1/4\alpha\log k-1/k)^{p^*}}\le\kappa_1\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}},
\end{aligned}
$$

where the second last inequality is due to Lemmas 9 and 7.

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

$$
\begin{aligned}
&\xi_r^k(t)=\frac{1}{m_k}\bar{E}_r^{o,k}(m_kt)-\lambda_r^kt,\\
&\tau_r^k(t)=\inf\{u\ge 0:|\xi_r^k(t+u)-\xi_r^k(t)|>2/\alpha\log k,\ u \text{ is a jump time}\}.
\end{aligned}\tag{70}
$$

We can write the event:

$$\begin{aligned}&\left\{\sup_{0\le u\le u_1^*}|\frac{1}{m_k}(\bar{E}_r^{o,k}(m_k(t+u))-\bar{E}_r^{o,k}(m_kt))-\lambda_r^k u|>\frac{3}{\alpha\log k}\right\}\cap\Omega_u^k\\ =&\left\{\sup_{0\le u\le u_1^*}|\xi_r^k(t+u)-\xi_r^k(t)|>\frac{3}{\alpha\log k}\right\}\cap\Omega_u^k\subset\{\tau_r^k(t)\le u_1^*\}\cap\Omega_u^k\end{aligned}\tag{71}$$

To see the inclusion in the above, consider any sample in the first (and hence the second) event. For this sample, we can find $u'\in[0,u_1^*]$ such that

$$|\xi_r^k(t+u')-\xi_r^k(t)|>\frac{3}{\alpha\log k}.$$

Let $u''(\le u')$ be the maximum jump time of $\xi_r^k(t+\cdot)$ before the time u'. Observe that the time length $(u'-u'')$ is a portion of a (scaled) interarrival time involved in the definition of $u_r^{k,\max}(k^2m_kt^*)$, and according to the definition of Ω_u^k, it must satisfy

$$u'-u''\le\frac{1}{km_k}u_r^{k,\max}(k^2m_kt^*)\le\frac{1}{k^{(p^*-2)/2p^*}}.$$

Moreover, as u'' is the only jump time in $[u'',u']$, we have for sufficient large k,

$$|\xi_r^k(t+u')-\xi_r^k(t+u'')|=\frac{1}{km_k}+\lambda_r^k(u'-u'').$$

The above three estimates together imply

$$|\xi_r^k(t+u'')-\xi_r^k(t)|\ge|\xi_r^k(t+u')-\xi_r^k(t)|-|\xi_r^k(t+u')-\xi_r^k(t+u'')|>\frac{2}{\alpha\log k},$$

which yields $\tau_r^k(t)\le u''\le u_1^*$.

Evaluate the following,

$$\begin{aligned}K:=&\,\mathsf{P}\left\{\tau_r^k(t)\le u_1^*,|\xi_r^k(t+u_1^*)-\xi_r^k(t+\tau_r^k(t)|\le\frac{1}{2\alpha\log k}\right\}\\ =&\int_{u\in[0,u_1^*]}F_{\tau_r^k(t)}(du)\mathsf{P}\left(|\xi_r^k(t+u_1^*)-\xi_r^k(t+u)|\le\frac{1}{2\alpha\log k}\;\Big|\;\tau_r^k(t)=u\right),\end{aligned}\tag{72}$$

where we denote the distribution $F_{\tau_r^k(t)}(u):=\mathsf{P}\left\{\tau_r^k(t)\le u\right\}$. For the integrand in the above, we have for sufficiently large k,

$$\begin{aligned}&\mathsf{P}\left(|\xi_r^k(t+u_1^*)-\xi_r^k(t+u)|\le\frac{1}{2\alpha\log k}\;\Big|\;\tau_r^k(t)=u\right)\\ =&\,\mathsf{P}\left(|\xi_r^k(u_1^*-u)-\xi_r^k(0)|\le\frac{1}{2\alpha\log k}\right)\\ \ge&\,1-\mathsf{P}\left(\left\{|\xi_r^k(u_1^*-u)-\xi_r^k(0)|>\frac{1}{2\alpha\log k}\right\}\cap\Omega_u^k\right)-\mathsf{P}(\Omega\setminus\Omega_u^k)\\ \ge&\,1-\kappa_1\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}-\frac{\kappa'}{k^{p^*/2-1}}\ge1-\kappa_2\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}},\end{aligned}$$

where the equality is because the (scaled and centered) renewal process $(\xi_r^k(t+\cdot)-\xi_r^k(t))$ restarts (probabilistically) at the stopping time $\tau_r^k(t)$, and the second inequality is due to (68, 69). The probability, K, can now be bounded from below:

$$K\ge\mathsf{P}\left\{\tau_r^k(t)\le u_1^*\right\}\left(1-\kappa_2\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}\right),$$

On the other hand, we can use (68, 69) again to estimate the upper bound of K as follows,

$$K \le \mathsf{P}\left\{|\xi_r^k(t+u_1^*)-\xi_r^k(t)|>\frac{1}{2\alpha\log k}\right\}\le \kappa_2\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}.$$

The above two bounds of K imply,

$$\mathsf{P}\left\{\tau_r^k(t)\le u_1^*,\omega\in\Omega_u^k\right\}\le\kappa_2\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}\left(1-\kappa_2\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}\right)^{-1}\le\kappa_3\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}},$$

for sufficiently large k.

Using the inclusion relationship in (71) and the above result inequlity, we bound the following probability,

$$\begin{aligned}
&\mathsf{P}\left\{\sup_{0\le t\le kt^*}\sup_{0\le u\le u^*}|\frac{1}{m_k}(\bar E_r^{o,k}(m_k(t+u))-\bar E_r^{o,k}(m_kt))-\lambda_r^k u|>\frac{6}{\alpha\log k},\omega\in\Omega_u^k\right\}\\
&=\mathsf{P}\left(\bigcup_{j=0}^{k-1}\left\{\sup_{jt^*\le t\le(j+1)t^*}\sup_{0\le u\le u^*}|\frac{1}{m_k}(\bar E_r^{o,k}(m_k(t+u))-\bar E_r^{o,k}(m_kt))-\lambda_r^k u|>\frac{6}{\alpha\log k},\omega\in\Omega_u^k\right\}\right)\\
&\le\sum_{j=0}^{k-1}\mathsf{P}\left\{\sup_{jt^*\le t\le(j+1)t^*}\sup_{0\le u\le u^*}|\frac{1}{m_k}(\bar E_r^{o,k}(m_k(t+u))-\bar E_r^{o,k}(m_kjt^*))-\lambda_r^k(t+u-jt^*)|\right.\\
&\quad\left.+|\frac{1}{m_k}(\bar E_r^{o,k}(m_kt)-\bar E_r^{o,k}(m_kjt^*))-\lambda_r^k(t-jt^*)|>\frac{6}{\alpha\log k},\omega\in\Omega_u^k\right\}\\
&\le\sum_{j=0}^{k-1}\mathsf{P}\left\{2\sup_{0\le u\le u_1^*}|\frac{1}{m_k}(\bar E_r^{o,k}(m_kjt^*+m_ku)-\bar E_r^{o,k}(m_kjt^*))-\lambda_r^k u|>\frac{6}{\alpha\log k},\omega\in\Omega_u^k\right\}\\
&\le\sum_{j=0}^{k-1}\mathsf{P}\left\{\tau_r^k(jt^*)\le u_1^*\right\}\le k\kappa_3\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-1}}\le\kappa_3\frac{(\alpha\log k)^{p^*}}{k^{p^*/2-2}}.
\end{aligned}$$

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

$$\mathsf{P}\left(\left(\Omega\setminus\Omega_E^k\right)\bigcap\Omega_u^k\right)\le\kappa_4\frac{(\log k)^{p^*}}{k^{p^*/2-2}}.$$

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

$$\mathsf{P}\left(\Omega\setminus\Omega_E^k\right)\le\kappa_5\frac{(\log k)^{p^*}}{k^{p^*/2-2}}.$$

5.1. 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, $i=1,2,\cdots$, 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 7 For any $r>0$, there exists a constant $a_r>0$ (depending on r and the distribution of X_1 only) such that

$$\mathsf{E}Y_t^r\le a_r(1+t^r),\quad t\ge 0. \tag{73}$$

The lemma is a direct result of the strong law of counting (renewal) process (e.g., Theorem 5.1 in Chapter 2 of Gut [11]), and hence its proof is omitted. Note that the lemma requires the existence of the first moment of X_i only.

Lemma 8 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

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

The first inequality can be found from Gut ([11], page 169), and the second follows from L^p maximum inequality (e.g., Theorem 5.4.3 of Durrett [8]).

Lemma 9 Let $r > p \geq 2$, and assume $\mathsf{E}X_i^r < \infty$. Then, there exists a constant c such that for all $t \geq 0$,

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

This lemma, to the best of our knowledge, first appeared as Theorem 4 of Krichagina and Taksar [15]), with a long proof. Here for completeness, we show it follows rather quickly from Lemmas 7 and 8.

Proof. Note that

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

and

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

Hence, we have

$$\sup_{0\leq s\leq t} |Y_s - \mu^{-1}s| \leq 3\mu^{-1} \sup_{0\leq s\leq t} |S_{Y_s+1} - \mu(Y_s+1)|. \tag{74}$$

Applying Lemma 8, we have

$$\begin{aligned} &\mathsf{E}\left(\sup_{0\leq s\leq t} |S_{Y_s+1} - \mu(Y_s+1)|^p \cdot 1_{\{Y_t < 2\mu^{-1}t+2\}}\right) \\ &\leq \mathsf{E}\left(\sup_{0\leq i\leq 2\mu^{-1}t+3} |S_i - \mu i|^p\right) \leq c_1'\left(1 + t^{\frac{p}{2}}\right). \end{aligned} \tag{75}$$

From (74) and (75), we have

$$\mathsf{E}\left(\sup_{0\leq s\leq t} |Y_s - \mu^{-1}s|^p \cdot 1_{\{Y_t < 2\mu^{-1}t+2\}}\right) \leq c_1\left(1 + t^{\frac{p}{2}}\right). \tag{76}$$

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

$$\begin{aligned} &\mathsf{E}\left(\sup_{0\leq s\leq t} |Y_s - \mu^{-1}s|^p \cdot 1_{\{Y_t \geq 2\mu^{-1}t+2\}}\right) \leq \mathsf{E}\left((Y_t^p + (\mu^{-1}t)^p) \cdot 1_{\{Y_t \geq 2\mu^{-1}t+2\}}\right) \\ &\leq \left(\mathsf{E}(Y_t^p + (\mu^{-1}t)^p)^\alpha\right)^{\frac{1}{\alpha}} \left(\mathsf{E}\left(1_{\{Y_t \geq 2\mu^{-1}t+2\}}\right)^\beta\right)^{\frac{1}{\beta}} \leq (c_2'(1 + t^{\alpha p}))^{\frac{1}{\alpha}} \left(\mathsf{P}\{Y_t \geq 2\mu^{-1}t+2\}\right)^{\frac{1}{\beta}} \\ &\leq c_2''(1+t^p)(\mathsf{P}\{S_{\lfloor 2\mu^{-1}t+2\rfloor} \leq t\})^{\frac{1}{\beta}} \leq c_2''(1+t^p)(\mathsf{P}\{|S_{\lfloor 2\mu^{-1}t+2\rfloor} - \mu\lfloor 2\mu^{-1}t+2\rfloor| \geq \mu + t\})^{\frac{1}{\beta}} \\ &\leq c_2''(1+t^p)\left(\frac{b_r \mathsf{E}|X_1 - \mu|^r \lfloor 2\mu^{-1}t+2\rfloor^{r/2}}{(\mu+t)^r}\right)^{\frac{1}{\beta}} \leq c_2\left(1 + t^{\frac{p}{2}}\right). \end{aligned} \tag{77}$$

Here we have applied Lemmas 7 and 8 in the third and the sixth inequalities respectively. Finally, the desired result is implied by the inequalities in (76) and (77). □

We remark that rigorously speaking, the inequalities in (44, 45) 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.

6. Appendix: Proof of Proposition 4 We return to prove Proposition 4. As a preparation, the next lemma claims that the fluid-scaled pre-limit networks $\bar{W}^{k,j}(u)$ can be approximated by the so-called fluid model, $(\bar{w}(t), \bar{u}(1), \bar{v}(1))$, which satisfies the following set of equations and conditions:

$$\begin{aligned}\bar{w}(t) &= \bar{w}(0) + \mathrm{diag}(\rho)(te - \bar{u}(1))^+ - (\bar{d}(t) - \bar{v}(1))^+ \\ &= \bar{w}(0) - \mathrm{diag}(\rho)(te \wedge \bar{u}(1)) + (\bar{d}(t) \wedge \bar{v}(1)) + \rho t - \bar{d}(t),\end{aligned} \tag{78}$$

$$\bar{d}_r(t) = \int_0^t \bar{\Lambda}_r(\bar{n}(s))ds, \tag{79}$$

$$\bar{\Lambda}_r(n) = \begin{cases} \Lambda_r(n) & \text{if } n_r > 0, \\ \rho_r & \text{if } n_r = 0. \end{cases} \tag{80}$$

The R-dimensional nonnegative vector function, $\bar{w}(t) \equiv \mathrm{diag}(\nu)\bar{n}(t)$, is interpreted as the fluid level process. The "residuals", $\bar{u}(1)$ and $\bar{v}(1)$, are R-dimensional vector constants, where the parameter "(1)" is attached in order to align the notation with its counterpart in the pre-limit networks. Here, we denote as e the vector with all components being one's. The dimension of e can be understood from the context.

Lemma 10 (Uniform Continuity) Let M be any given positive numbers (and Δ, T, $\bar{\Delta}$ and $\bar{T}$ are specified as in Proposition 4).
(a) For any $\epsilon > 0$, there exists k^* such that for any $k \geq k^*$, the following holds for any $m_k \geq |\hat{\Xi}^k(0)| \vee 1$, $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, and $0 \leq j \leq k\Delta/y^k T$: if

$$|\bar{W}^{k,j}(0)| + |\bar{U}^{k,j}(0)| + |\bar{V}^{k,j}(0)| \leq M, \tag{81}$$

then, we can find a fluid model $(\bar{w}(t), \bar{u}(1), \bar{v}(1))$ satisfying (78-80) and $|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq M$ such that

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

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

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

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

$$|\bar{W}^{k,j}(0)| \leq M, \tag{82}$$

then, we can find a fluid model $(\bar{w}(t), \bar{u}(1) = 0, \bar{v}(1) = 0)$ satisfying (78-80) and $|\bar{w}(0)| \leq M$ such that

$$\sup_{0 \leq u \leq T} |\bar{W}^{k,j}(u) - \bar{w}(u)| < \epsilon.$$

The proof of the above lemma, along with other preliminary results needed (e.g., the uniform attraction, complementarity and oscillation inequality), is deferred to subsections §6.1 and §6.2 below.

The proof of Proposition 4 is a modification of the proof of Lemma 7 of [19] by carefully accommodating the extra scaling factor y^k and the regular events.

We specify the time length of T as follows:

$$T \geq \max\{T_{\kappa,\epsilon/8}, T_{\kappa,\sigma/2}\}, \tag{83}$$

where the terms on the right hand side are defined in Lemma 12, and $\sigma = \sigma(\kappa', \epsilon/2)$ is specified in Lemma 13. Note that T is large enough so that in the fluid network in Lemma 12 (under the heavy traffic condition), the state $\bar{w}(t)$ will be close enough (by an error bound of $\epsilon/8$ or $\sigma/2$) to the fixed-point state, starting from an initial state $(\bar{w}(0), \bar{u}(1), \bar{v}(1))$ that is bounded by κ. Here, ϵ is given in the current lemma under proof, and κ and κ' are constants that depend on network parameters only:

$$\kappa = \kappa_w + 2\kappa_c + 1, \quad \kappa' = \kappa_w \cdot \kappa + 1, \tag{84}$$

where κ_w and κ_c are given in Lemma 12 and Lemma 14. The rationale for the choice of both κ and κ' will be clear in the context of the proof.

Step 1. We prove the three parts of Proposition 4, (a,b,c), for $j = 1$.

Let $\epsilon' > 0$ be any given number. Note that $|\bar{\Xi}^{k,0}(0)| = |\hat{\Xi}^k(0)/y^k| \leq 1$ according to the definitions in (58, 59). By Lemma 10, we have for sufficiently large k, and for any initial state $\hat{\Xi}^k(0)$, $m_k \geq (|\hat{\Xi}^k(0)| \vee 1)$ and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, there exists a fluid model $(\bar{w}(u), \bar{u}(1), \bar{v}(1))$ satisfying (78-80), which may depend on k, $\hat{\Xi}^k(0)$, m_k and ω, such that,

$$\begin{array}{l} |\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq 1 \ (\leq \kappa), \quad \text{and} \\ \sup\limits_{u\in[0,2T]} \big(|\bar{W}^{k,0}(u) - \bar{w}(u)| + |\bar{U}^{k,0}(0) - \bar{u}(1)| + |\bar{V}^{k,0}(0) - \bar{v}(1)|\big) < \epsilon'. \end{array} \tag{85}$$

Since $T \geq T_{\kappa,\epsilon/8}$, applying the uniform attraction property in Lemma 12 to the above $\bar{w}(u)$ yields:

$$d^{fp}(\bar{w}(u)) \leq \frac{\epsilon}{8} \ \text{ for all } u \geq T; \ \text{ and } \ |\bar{w}(u)| \leq \kappa_w(|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)|) \ \text{ for all } u \geq 0. \tag{86}$$

Note that $\bar{W}^{k,0}(T+u) \equiv \bar{W}^{k,1}(u)$ (and $\bar{W}^{k,0}(0) = \hat{W}^k(0)/y^k$). Hence, choosing a sufficiently small ϵ' at the beginning of the proof, the estimate in (85), along with (86), implies that the conclusion (a) holds with $j = 1$ for sufficiently large k and for all $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$. By (85, 86) again, we have for all $u \in [0, 2T]$,

$$|\bar{W}^{k,0}(u)| \leq |\bar{w}(u)| + \epsilon' \leq \kappa_w |\bar{\Xi}^{k,0}(0)| + 2\epsilon' \leq \kappa_w + \epsilon,$$

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

$$|\bar{W}^{k,1}(u)| = |\bar{W}^{k,0}(T+u)| \leq \kappa_w + \epsilon \ (\leq \kappa \wedge \kappa'). \tag{87}$$

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

Furthermore, since $T \geq T_{\kappa,\sigma/2}$, the first inequality in (86) also hold with $\epsilon/8$ replaced by $\sigma/2$, i.e., $d^{fp}(\bar{w}(u)) \leq \sigma/2$ for $u \geq T$; and therefore, the result in (a), with ϵ replaced by σ as well, holds with $j = 1$, i.e., for any sufficiently large k and for any $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$,

$$d^{fp}(\bar{W}^{k,1}(u)) \leq \sigma, \quad \text{for } u \in [0, T]. \tag{88}$$

In addition to (86), we can require the following via Lemma 10 and Lemma 12 too:

$$|G^T(\bar{w}(T+u)-w^*)| \le \frac{\epsilon}{8}, \tag{89}$$

$$|G^T(\bar{W}^{k,1}(u)-\bar{w}(T+u))| = |G^T(\bar{W}^{k,0}(T+u)-\bar{w}(T+u))| \le \frac{\epsilon}{8}, \tag{90}$$

for $u \in [0,T]$ and for some fixed-point state w^* (which is associated with k and ω too). Now, consider any server ℓ satisfying the "if" condition in (b) for $j=1$. Using the estimations in (89) and (90), we have further that for any $u \in [0,T]$,

$$\begin{aligned} & |g_\ell^T(\bar{W}^{k,1}(u)-\bar{W}^{k,1}(u'))| \\ \le\ & |g_\ell^T(\bar{W}^{k,1}(u)-\bar{w}(T+u))| + |g_\ell^T(\bar{w}(T+u)-w^*)| \\ & + |g_\ell^T(w^*-\bar{w}(T+u'))| + |g_\ell^T(\bar{w}(T+u')-\bar{W}^{k,1}(u'))| \\ \le\ & \frac{\epsilon}{8}+\frac{\epsilon}{8}+\frac{\epsilon}{8}+\frac{\epsilon}{8} = \frac{\epsilon}{2}, \end{aligned}$$

and hence

$$g_\ell^T \bar{W}^{k,1}(u) \ge g_\ell^T \bar{W}^{k,1}(u') - \frac{\epsilon}{2} \ge \frac{\epsilon}{2}. \tag{91}$$

Thereafter, we have

$$\bar{Y}_\ell^{k,1}(u) - \bar{Y}_\ell^{k,1}(0) = \int_0^u \left(c_\ell - A_\ell \Lambda(\bar{W}^{k,1}(s))\right) ds = 0, \tag{92}$$

where the first equality follows from the definitions of the processes $\bar{Y}^{k,j}(u)$ and $\hat{Y}^k(t)$; and in the second equality we have applied Lemma 13 to the server ℓ given the upper bound in (87) and the estimations in (88) and (91).

Step 2. We now extend the above to $j = 2, \ldots, k\delta/y^kT$. 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,c) does not hold for some integer $j \in [2, k\delta/y^kT]$ and for some $\hat{\Xi}^k(0)$, $m_k \ge (|\hat{\Xi}^k(0)| \vee 1)$ and sample-path $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$. Let j_k be the smallest positive integer in the interval $[2, k\delta/y^kT]$ such that at least one of the properties in (a, b, c) 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, c) 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 $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$.

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

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

By Lemma 10(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 $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, there exists a fluid model $\bar{w}(u)$ satisfying (78-80) (which may depend on k, $\hat{\Xi}^k(0)$, m_k and ω) such that

$$\sup_{u \in [0,2T]} |\bar{W}^{k,j_k-1}(u) - \bar{w}(u)| < \epsilon' \tag{93}$$

with $|\bar{w}(0)| \le \kappa$. (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 10(b).) Since $T \ge T_{\kappa,\epsilon/8}$, applying the uniform attraction property in Lemma 12 to the above limit yields:

$$d^{fp}(\bar{w}(u)) \le \frac{\epsilon}{8} \ \text{ for all } u \ge T. \tag{94}$$

Note that $\bar{W}^{k,j_k-1}(T+u) \equiv \bar{W}^{k,j_k}(u)$. Hence, the convergence in (93), along with (94), implies that (a) holds with $j = j_k$ and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, for sufficiently large $k \in \mathcal{K}_1$.

In addition to (94), we can require the following via Lemma 10 and Lemma 12 too:

$$|\bar{w}(u)| \le \kappa_w |\bar{w}(0)| \le \kappa_w \kappa, \quad \text{for all } u \ge 0; \tag{95}$$

$$|G^T(\bar{w}(T+u) - w^*)| \le \frac{\epsilon}{8}, \quad \text{for } u \ge 0, \text{ and for some } w^* \in \mathcal{W}; \tag{96}$$

$$|G^T(\bar{W}^{k,j_k}(u) - \bar{w}(T+u))| = |G^T(\bar{W}^{k,j_k-1}(T+u) - \bar{w}(T+u))| \le \frac{\epsilon}{8}, \tag{97}$$
$$\text{for } u \in [0,T], \text{ and for sufficiently large } k \in \mathcal{K}_1.$$

(Keep in mind that the above hold for all $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, and that $\bar{w}(u)$ and w^* are associated with k and ω too.) The first bound above implies the following for sufficiently large $k \in \mathcal{K}_1$, for $u \in [0,T]$ and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$:

$$|\bar{W}^{k,j_k}(u)| \le |\bar{w}(T+u)| + \epsilon \le \kappa_w |\bar{w}(0)| + \epsilon \le \kappa_w \kappa + \epsilon = \kappa'. \tag{98}$$

Furthermore, since $T \ge T_{\kappa,\sigma/2}$, the inequality in (94) also hold with $\epsilon/8$ replaced by $\sigma/2$, i.e., $d^{fp}(\bar{w}(u)) \le \sigma/2$ and therefore, the result in (a), with ϵ replaced by σ as well, holds with $j = j_k$ and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, for sufficiently large $k \in \mathcal{K}_1$, i.e.,

$$d^{fp}(\bar{W}^{k,j_k}(u)) \le \sigma, \quad \text{for } u \in [0,T]. \tag{99}$$

Now, consider any server ℓ and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$ satisfying the "if" condition in (b) for $j = j_k$. Similar to (91) and (92), we use the estimations in (96) and (97) to show that for sufficiently large $k (\in \mathcal{K}_1)$ and for any $u \in [0,T]$ and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$,

$$g_\ell^T \bar{W}^{k,j_k}(u) \ge \bar{W}^{k,j_k}(u') - \frac{\epsilon}{2} \ge \frac{\epsilon}{2}, \tag{100}$$

and thereafter apply the estimations in (98, 99, 100) to derive the following,

$$\bar{Y}_\ell^{k,j_k}(u) - \bar{Y}_\ell^{k,j_k}(0) = \int_0^u \left(c_\ell - A_\ell \Lambda(\bar{N}^{k,j_k}(s))\right) ds = 0. \tag{101}$$

Consider any sufficiently large $k \in \mathcal{K}_1$, such that the results in (a) and (b) hold for $j = 1, \cdots, j_k$ (but (a) needs not holds for $j = 0$) and for all $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$. Fix any $\hat{\Xi}^k(0)$, $m_k \ge (|\hat{\Xi}^k(0)| \vee 1)$, and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$. 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 14 *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 $j = 1, \cdots, j_k$. Hence, we have for any $t \in [y^k T/k, (j_k y^k T + y^k T)/k]$ $(\subset [0, \Delta])$,

$$\begin{aligned} &\mathsf{Osc}\left(\frac{1}{y^k m_k} \hat{W}^k(m_k s), s \in [y^k T/k, t]\right) \\ &\le \kappa_c \left(\mathsf{Osc}\left(\frac{1}{y^k m_k} \hat{X}^k(m_k s), s \in [y^k T/k, t]\right) + \epsilon\right) = \kappa_c(2+\epsilon). \end{aligned} \tag{102}$$

Consequently, we have the following estimations,

$$\begin{aligned} \left|\frac{1}{y^k m_k} \hat{W}^k(m_k t)\right| &\le \left|\frac{1}{y^k m_k} \hat{W}^k(m_k y^k T/k)\right| + \mathsf{Osc}\left(\frac{1}{y^k m_k} \hat{W}^k(m_k s), s \in [y^k T/k, t]\right) \\ &\le \kappa_w + \epsilon + \kappa_c(2+\epsilon), \end{aligned}$$

where in the second inequality we have also applied the conclusion in (87), i.e., $|\hat{W}^k(m_k y^k T/k)/y^k m_k| = |\bar{W}^{k,1}(0)| \le \kappa_w + \epsilon$. 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 (c) holds with $j = j_k$ for sufficiently large $k \in \mathcal{K}_1$.

6.1. Proof of Lemma 10 To prove the lemma, we first need some preliminary results.

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

$$w(t) = w(0) - \text{diag}(\rho)(te \wedge u(1)) + (d(t) \wedge v(1)) + x(t) + \rho t - d(t) \geq 0, \tag{103}$$

$$d_r(t) = \int_0^t \Lambda_r(n(s))ds. \tag{104}$$

In the above, $w(t)$ $(= \text{diag}(\nu)n(t))$ is an R-dimensional nonnegative vector function of time $t \geq 0$, which can be interpreted as the (generic and scaled) workload process. $x(t)$ is also an R-dimensional vector function of time $t \geq 0$, associated with the "free process" in the pre-limit 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 (78-80). 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 11(Uniform Continuity) Let T and M be any given positive numbers. For any ϵ, there exists a $\delta > 0$ such that for any $(w(t), x(t), u(1), v(1))$ satisfying (103-104) and

$$|w(0)| + |u(1)| + |v(1)| \leq M, \quad \sup_{0 \leq t \leq T} |x(t)| < \delta, \tag{105}$$

we can find a fluid model $(\bar{w}(t), \bar{u}(1), \bar{v}(1))$ satisfying (78-80) and

$$|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq M, \quad \sup_{0 \leq t \leq T} |w(t) - \bar{w}(t)| + |u(1) - \bar{u}(1)| + |v(1) - \bar{v}(1)| < \epsilon.$$

In addition, if the condition in (105) 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.

Proof. If to the contrary, we can find an $\epsilon_0 > 0$ and a sequence $\{(w^{(i)}(t), x^{(i)}(t), u^{(i)}(1), v^{(i)}(1)); i = 1, 2, \cdots\}$ satisfying (103-104) and

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

such that for all i, the following holds,

$$\sup_{0 \leq t \leq T} |w^{(i)}(t) - \bar{w}(t)| + |u^{(i)}(1) - \bar{u}(1)| + |v^{(i)}(1) - \bar{v}(1)| \geq \epsilon_0, \tag{106}$$

for any fluid model $(\bar{w}(t), \bar{u}(1), \bar{v}(1))$ satisfying (78-80) with $|w(0)| + |u(1)| + |v(1)| \leq M$.

Applying the convenitional approach for proving fluid limit theorem, however, we can find a subsequence of i such that as $i \to \infty$ along the subsequence, we have

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

for some fluid model $(\bar{w}(t), \bar{u}(1), \bar{v}(1))$ satisfying (78-80) with $|w(0)| + |u(1)| + |v(1)| \leq M$, which contradicts to (106).

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

To apply the above lemma for proving Lemma 10, we first spell out the the dynamics of the networks $\bar{W}^{k,j}(u)$ in period $[0,T]$ (i.e., $\hat{W}^k(m_k t)/y^k m_k$ in $[jy^kT/k,(j+1)y^kT/k]$ or $W^k(t)/ky^km_k$ in $[jky^km_kT,(j+1)ky^km_kT]$). The original and unscaled arrival process, restarted at time jky^km_kT, is also a (delayed) renewal process, which we denote as,

$$E_r^{k,j}(t) = E_r^k(jky^km_kT+t) - E_r^k(jky^km_kT).$$

It is defined by the delayed starting time $U_r^{k,j}(0) := U_r^k(jky^km_kT)$ (the "initial" residual arrival time) and the renewal sequence $\{u_r^k(i) : i \geq E_r^k(jky^km_kT)+2\}$. Denote the corresponding non-delayed version as $E_r^{o,k,j}(t)$, which is then defined by the renewal sequence $\{u_r^k(i) : i \geq E^k(jky^km_kT)+2\}$. Denote $\bar{U}_r^{k,j}(0) = U_r^{k,j}(0)/ky^km_k$, $\bar{E}_r^{k,j}(u) = E_r^{k,j}(ky^km_ku)/ky^km_k$ and $\bar{E}_r^{o,k,j}(u) = E_r^{o,k,j}(ky^km_ku)/ky^km_k$. Then,

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

The service process is characterized as,

$$\begin{aligned}S_r^{k,j}(t) &= S_r^k(D_r(jky^km_kT)+t) - S_r^k(D_r(jky^km_kT)),\\ D_r^{k,j}(t) &= D_r^k(jky^km_kT+t) - D_r^k(jky^km_kT) = \int_{jky^km_kT}^{jky^km_kT+t} \Lambda_r(N^k(s))ds.\end{aligned}$$

That is, $S_r^{k,j}(t)$ is defined by the delayed starting time $V_r^{k,j}(0) := V_r^k(D_r^k(jky^km_kT))$ (the "initial" residual arrival time) and the renewal sequence $\{v_r^k(i) : i \geq S_r^k(D_r(jky^km_kT))+2\}$. Denote the corresponding non-delayed version as $S_r^{o,k,j}(t)$, which is then defined by the renewal sequence $\{v_r^k(i) : i \geq S_r^k(D_r(jky^km_kT))+2\}$. Denote $\bar{V}_r^{k,j}(0) = V_r^{k,j}(0)/ky^km_k$, $\bar{S}_r^{k,j}(u) = S_r^{k,j}(ky^km_ku)/ky^km_k$, $\bar{S}_r^{o,k,j}(u) = S_r^{o,k,j}(ky^km_ku)/ky^km_k$, and $\bar{D}_r^{k,j}(u) = D_r^{k,j}(ky^km_ku)/ky^km_k$. Then, we write

$$\begin{aligned}\bar{S}_r^{k,j}(u) - (\nu_r^k)^{-1}u = {} & \bar{S}_r^{o,k,j}([u-\bar{V}_r^{k,j}(0)]^+) - (\nu_r^k)^{-1}[u-\bar{V}_r^{k,j}(0)]^+ \\ & +1_{\{u\geq \bar{V}_r^{k,j}(0)\}}/ky^km_k - (\nu_r^k)^{-1}(u\wedge \bar{V}_r^{k,j}(0)),\end{aligned}$$

and

$$\bar{D}_r^{k,j}(u) = \int_0^u \Lambda_r(\bar{N}^{k,j}(s))ds. \tag{107}$$

Finally, the workload process can be written as,

$$\begin{aligned}\bar{W}_r^{k,j}(u) &= \bar{W}_r^{k,j}(0) + \nu_r^k\bar{E}_r^{k,j}(u) - \nu_r^k\bar{S}_r^{k,j}(\bar{D}_r^{k,j}(u)) \\ &= \bar{W}_r^{k,j}(0) + \nu_r^k\left(\bar{E}_r^{k,j}(u) - \lambda_r^k u\right) - \left(\nu_r^k\bar{S}_r^{k,j}(\bar{D}_r^{k,j}(u)) - \bar{D}_r^{k,j}(u)\right) \\ &\quad + \left(\rho_r^k u - \rho_r u\right) + \left(\rho_r u - \bar{D}_r^{k,j}(u)\right) \\ &= \bar{W}^{k,j}(0) - \rho_r(u\wedge\bar{U}_r^{k,j}(0)) + (\bar{D}_r^{k,j}(u)\wedge\bar{V}_r^{k,j}(0)) \\ &\quad +\nu_r^k\left(\bar{E}_r^{o,k,j}([u-\bar{U}_r^{k,j}(0)]^+) - \lambda_r^k[u-\bar{U}_r^{k,j}(0)]^+\right) \\ &\quad -\left(\nu_r^k\bar{S}_r^{o,k,j}([\bar{D}_r^{k,j}(u)-\bar{V}_r^{k,j}(0)]^+) - [\bar{D}_r^{k,j}(u)-\bar{V}_r^{k,j}(0)]^+\right) \\ &\quad -\frac{\nu_r^k}{ky^km_k}\left(1_{\{u\geq\bar{U}_r^{k,j}(0)\}} + 1_{\{\bar{D}_r^{k,j}(u)\geq\bar{V}_r^{k,j}(0)\}}\right) \\ &\quad -(\rho_r^k-\rho_r)(u\wedge\bar{U}_r^{k,j}(0)) + \left(\rho_r^k-\rho_r\right)u + \left(\rho_r u - \bar{D}_r^{k,j}(u)\right).\end{aligned} \tag{108}$$

Proof (of Lemma 10). To apply Lemma 11, we denote the terms in (108) which we show will vanish as follows for convenience,

$$\begin{aligned} x_r(u) := \nu_r^k \Sigma_E - \nu_r^k \Sigma_S + (\rho_r^k - \rho_r)(u - u \wedge \bar{U}_r^{k,j}(0)) \\ + \frac{\nu_r^k}{k y^k m_k} \left(1_{\{u \geq \bar{U}_r^{k,j}(0)\}} - 1_{\{\bar{D}_r^{k,j}(u) \geq \bar{V}_r^{k,j}(0)\}} \right), \end{aligned} \tag{109}$$

where

$$\begin{aligned} \Sigma_E &:= \bar{E}_r^{o,k,j}([u - \bar{U}_r^{k,j}(0)]^+) - \lambda_r^k [u - \bar{U}_r^{k,j}(0)]^+, \\ \Sigma_S &:= \bar{S}_r^{o,k,j}([\bar{D}_r^{k,j}(u) - \bar{V}_r^{k,j}(0)]^+) - \mu_r^k [\bar{D}_r^{k,j}(u) - \bar{V}_r^{k,j}(0)]^+. \end{aligned}$$

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

$$\begin{aligned} |\Sigma_E| &= \frac{1}{k y^k m_k} |E_r^{o,k}(k m_k \tau + k y^k m_k [u - \bar{U}_r^{k,j}(0)]^+) \\ &\quad - E_r^{o,k}(k m_k \tau) - k y^k m_k \lambda_r^k [u - \bar{U}_r^{k,j}(0)]^+| \\ &\leq \sup_{u' \in [0,T]} \frac{1}{y^k} | \frac{1}{k m_k} (E_r^{o,k}(k m_k (\tau + y^k u')) - E_r^{o,k}(k m_k \tau)) - k y^k \lambda_r^k u'| \\ &= \frac{1}{y^k} \sup_{u' \in [0,T]} | \frac{1}{m_k} (\bar{E}_r^{o,k}(m_k (\tau + y^k u')) - \bar{E}_r^{o,k}(m_k \tau)) - y^k \lambda_r^k u'| \\ &\leq \frac{1}{y^k} \sum_{i=1}^{\lceil y^k \rceil} \sup_{u' \in [0,T]} | \frac{1}{m_k} (\bar{E}_r^{o,k}(m_k (\tau + (i-1)T + u')) \\ &\quad - \bar{E}_r^{o,k}(m_k (\tau + (i-1)T))) - \lambda_r^k u'| \end{aligned} \tag{110}$$

(Here we denote $\tau := j y^k T - U_r^{k,0}(0)/k m_k + U_r^{k,j}(0)/k m_k$ for convenience.) Estimate the time $\tau + (i-1)T$ inside the supremum above for $i \leq \lceil y^k \rceil$,

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

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

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

for sufficiently large k. The above inequality indicates that the time periods involved in (110) fall within those covered in $\Omega_E^k(\bar{\Delta}, \bar{T}, m_k)$, so that the bound in that event can be invoked for each item in (110); that is, we have for sufficiently large k, for any $m_k \geq (|\hat{\Xi}^k(0)| \vee 1)$ and $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$, for all $i = 1, \cdots, \lceil y^k \rceil$,

$$\begin{aligned} &\sup_{u' \in [0,T]} | \frac{1}{m_k} (\bar{E}_r^{o,k}(m_k (\tau + (i-1)T + u')) - \bar{E}_r^{o,k}(m_k (\tau + (i-1)T))) - \lambda_r^k u'| \\ \leq &\sup_{t \in [0, k\bar{\Delta}]} \sup_{u' \in [0,\bar{T}]} | \frac{1}{m_k} (\bar{E}_r^{o,k}(m_k (t + u')) - \bar{E}_r^{o,k}(m_k t)) - \lambda_r^k u'| \leq \delta, \end{aligned} \tag{111}$$

where the second inequality follows from the definition of $\Omega^k(\bar{\Delta}, \bar{T}, m_k)$ ($\subset \Omega_E^k(\bar{\Delta}, \bar{T}, m_k)$) with sufficiently large k (say, $1/\log k < \delta$). Then, we have from (110) and (111),

$$\Sigma_E \leq \frac{1}{y^k} \lceil y^k \rceil \delta \leq 2\delta. \tag{112}$$

Second, we estimate the term involving the service process, Σ_S. The approach is similar to the estimation of Σ_E. Denote $\bar{c} = \max\{c_\ell\} \vee 1$ for convenience. For any $u \in [0,T]$, we have

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

(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 $u \in [0,T]$,

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

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

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

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 $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k)$,

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

and thereafter,

$$\Sigma_S \le 2\delta. \tag{113}$$

From (109, 112, 113), we know that the condition in (105) in Lemma 11 (in particular, $\sup_{0\le t\le T}|x(t)| < \delta$) can be justified for all sufficiently large k, all $j = 0, 1, \cdots, k\Delta/y^k T$ and all $\omega \in \bar{\Omega}^k(\bar{\Delta}, \bar{T}, m_k)$. Therefore, Lemma 11 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

$$T > \bar{U}_r^{k,0}(0), \quad \bar{D}_r^{k,0}(T) > \bar{V}_r^{k,0}(0), \tag{114}$$

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 Lemma 12: given the bounded initial state $|\bar{\Xi}^{k,0}(0)| = |\hat{\Xi}^k(0)|/y^k m_k \le 1$, the process $\bar{D}_r^{k,0}(t)$ is close to $\bar{d}_r(t)$ for sufficiently large k, whereas $\bar{d}_r(t)$ is close to $\rho_r t$ for sufficiently large t. The inequalities in (114) implies that for $j \ge 1$, $U_r^{k,j}(0)$ (resp. $V_r^{k,j}(0)$) must be a portion of an interarrival time (resp. a service requirement) of class-r *other than* the initial residual arrival time $u_r^k(1)$ (resp. initial service requirement $v_r^k(1)$) of the original k-th network. Hence, following the definition in (50, 51), we have for $1 \le j \le k\bar{\Delta}/y^k T$ and $r \in \mathcal{R}$,

$$U_r^{k,j}(0) \le u_r^{k,\max}(k^2 m_k \bar{\Delta}), \quad V_r^{k,j}(0) \le v_r^{k,\max}(k^2 m_k \bar{\Delta}).$$

Consequently, give the assumption $\omega \in \Omega^k(\bar{\Delta}, \bar{T}, m_k) \subset \Omega_u^k(\bar{\Delta}, m_k) \cap \Omega_v^k(\bar{\Delta}, m_k)$, the above inequalities imply the first conclusion in part (b), which along with the last conclusion in Lemma 11, further implies the second conclusion in part (b). □

6.2. Uniform Attraction, Complementarity and Oscillation Inequality We replicate some useful tools from our previous work [20]; also refer to that paper for the original references.

Lemma 12 Consider the fluid limit $\bar{w}(t)$ in Lemma 10 (i.e., the fluid model in (78-80)), along with the constant M and τ specified there. Assume the heavy traffic condition in (12) holds.
(a) The server-based workload, $A_\ell \bar{w}(t)$ ($\ell \in \mathcal{L}$), is non-decreasing in time $t \geq M\tau$; and there exists a constant κ_w that only depends on the network parameters, such that the following bounds hold for all $t \geq 0$,

$$|\bar{w}(t)| \leq \kappa_w(|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)|) \ (\leq \kappa_w M). \tag{115}$$

(b) (Uniform Attraction) There exists a (unique) fixed-point state w^* such that, for any given $\epsilon > 0$ and for some sufficiently large time $T_{M,\epsilon}$ (depending on M and ϵ), the following holds:

$$|\bar{w}(t) - w^*| \leq \epsilon, \quad \text{for } t \geq T_{M,\epsilon}. \tag{116}$$

Furthermore, the time $T_{M,\epsilon}$ can be chosen large enough such that the following also holds:

$$d^{fp}(\bar{w}(t)) \leq |G^T(\bar{w}(t) - w^*)| + |H^T \bar{w}(t)| \leq \epsilon, \quad \text{for } t \geq T_{M,\epsilon}. \tag{117}$$

(c) If $\bar{w}(0)$ is a fixed-point state and $(\bar{u}(1), \bar{v}(1)) = 0$, then $\bar{w}(t) = \bar{w}(0)$ and $\bar{d}(t) = \rho t$ for all $t \geq 0$.

The following lemma characterizes the reflection property of the regulator $\hat{Y}^k(t)$ $(= k\int_0^t (c - A\Lambda(\hat{N}^k(s)))ds)$ given in (25).

Lemma 13(Complementarity) Let $\kappa > 0$ and $\epsilon > 0$ be given constants. Then, there exists a (sufficiently small) constant $\sigma > 0$ such that, for any state w satisfying

$$|w| \leq \kappa \ \text{ and } \ d^{fp}(w) \leq \sigma, \tag{118}$$

the following implication holds for any $\ell \in \mathcal{L}$:

$$g_\ell^T w > \epsilon \qquad \Rightarrow \qquad A_\ell \Lambda(n) = \sum_{r \in \mathcal{R}} a_{\ell r} \Lambda_r(n) = c_\ell. \tag{119}$$

In words, the server ℓ will be fully occupied if the workload state of the network is away from the ℓ-facet (corresponding to $g_\ell^T w = 0$), and toward the interior, of the fixed-point state space $\mathcal{W}$.

The oscillation inequality is a useful tool to establish the boundedness of the workload process (refer to Proposition 4(c)); refer to [12, 20] for the following form of the inequality. To state the inequality, denote for any RCLL (vector) function $f(u)$ ($u \geq 0$) and any time interval $[s,t]$,

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

Lemma 14(Oscillation Inequality) Suppose there exists a constant $\kappa_c > 0$ such that, for any $\epsilon \geq 0$ and any RCLL functions, $w(t) = (w_\ell(t))_{\ell \in \mathcal{L}}$, $x(t) = (x_r(t))_{r \in \mathcal{R}}$, $y(t) = (y_\ell(t))_{\ell \in \mathcal{L}}$ and $z(t) = (z_m(t))_{m=1}^{R-L}$, satisfying

$$\begin{aligned}
&w(t) = w(0) + x(t) + BGy(t) + BHz(t) \ (\geq 0), \quad \text{for } t \geq 0;\\
&G^T w(t) \geq -\epsilon, \quad \text{for } t \geq 0;\\
&y_\ell(t) \text{ is non-decreasing in } t \geq 0, \ y_\ell(0) = 0, \quad \ell \in \mathcal{L};\\
&y_\ell(t) \text{ can not increase at time } t, \text{ if } g_\ell^T w(t) \geq \epsilon.
\end{aligned}$$

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

$$\mathsf{Osc}(G^T w(\cdot), [s,t]) \text{ and } \mathsf{Osc}(y(\cdot), [s,t]) \le \kappa_c(\mathsf{Osc}(x(\cdot), [s,t]) + \epsilon). \tag{120}$$

If in addition,

$$|H^T w(t)| \le \epsilon, \quad \text{for } t \ge 0, \tag{121}$$

then the above oscillation inequalities can be strengthened as follows: for any $0 \le s \le t$,

$$\mathsf{Osc}(w(\cdot), [s,t]) \text{ and } \mathsf{Osc}(y(\cdot), [s,t]) \le \kappa_c(\mathsf{Osc}(x(\cdot), [s,t]) + \epsilon). \tag{122}$$

Acknowledgments. Heng-Qing Ye was supported in part by HK/RGC Grant 15508114. David Yao was supported in part by NSF Grant CMMI-1462495.

References

[1] Billingsley, P., *Convergence of Probability Measures* (2ed), John Wiley & Sons, New York, 1999.

[2] Bramson, M., State Space Collapse with Application to Heavy Traffic Limits for Multiclass Queueing Networks. *Queueing Systems, Theory and Applications*. **30** (1998), 89-148.

[3] Budhiraja, A. and C. Lee, Stationary Distribution Convergence for Generalized Jackson Networks in Heavy Traffic. *Mathematics of Operations Research*, **34** (2009), 1, 45-56.

[4] Dai, J.G., On Positive Harris Recurrence of Multi-class Queueing Networks: A Unified Approach via Fluid Limit Models. *Annals of Applied Probability*, **5** (1995), 49-77.

[5] Dai, J.G. and S.P. Meyn, Stability and Convergence of Moments for Multiclass Queueing Networks via Fluid Models. *IEEE Transactions on Automatic Control*, **40** (1995), 1899-1904.

[6] Davis M.H.A., Piecewise-Deterministic Markov Processes: A General Class of Nondiffusion Models. *Journal of the Royal Statistical Society, Series B*, **46** (1984), 353-388.

[7] Dupuis, P. and R.J. Williams, Lyaponov Functions for Semimartingale Reflected Brownian Motions. *Annals of Applied Probability*, **22** (1994), 680-702.

[8] Durrett, R., *Probability: Theory and Examples* (4ed), Cambridge University Press, Cambridge, United Kingdom, 2010.

[9] Gamarnik D. and A. Zeevi, Validity of Heavy Traffic Steady-State Approximations in Generalized Jackson Networks. *Annals of Applied Probability*, **16** (2006), 56-96.

[10] Gurvich, I., Validity of Heavy-Traffic Steady-State Approximations in Multiclass Queueing Networks: The Case of Queue-Ratio Disciplines. *Mathematics of Operations Research*, **39** (2014), 121-162.

[11] Gut, A., *Stopped Random Walks: Limit Theorems and Applications*, vol. 5 of *Applied probability*, Springer-Verlag, New York, 1988.

[12] Kang, W.N., F.P. Kelly, N.H. Lee and R.J. Williams, State Space Collapse and Diffusion Approximation for a Network Operating under a Fair Bandwidth Sharing Policy. *Annals of Applied Probability*, **19** (2009), 1719-1780.

[13] Katsuda, T., State-Space Collapse in Stationarity and Its Application to a Multiclass Single-Server Queue in Heavy Traffic. *Queueing Systems: Theory and Applications*, **65** (2010), 237-273.

[14] Katsuda, T., Stationary Distribution Convergence for A Multiclass Single-Server Queue in Heavy Traffic. Scientiae Mathematicae Japonicae, **75** (2012), 317334.

[15] Krichagina, E.V. and M.I. Taksar, Diffusion Approximation for GI/G/1 Controlled Queues. *Queueing Systems: Theory and Applications*, **12** (1992), 333-368.

[16] Shah, D., J.N. Tsitsiklis and Y. Zhong, Qualitative Properties of Alpha-Fair Policies in Bandwidth-Sharing Networks. *Annals of Applied Probability*, **24** (2014), No.1, 76-113.

[17] Wang, W., S.T. Maguluri, R. Srikant and L. Ying, Heavy-Traffic Insensitive Bounds for Weighted Proportionally Fair Bandwidth Sharing Policies. Available at https://arxiv.org/abs/1808.02120

[18] Williams, R.J., Diffusion Approximations for Open Multi-class Queueing Networks: Sufficient Conditions Involving State Space Collapse. *Queueing Systems, Theory and Applications*, **30** (1998), 27-88.

[19] Ye, H.Q. and D.D. Yao, A Stochastic Network under Fair Resource Control — Diffusion Limit with Multiple Bottlenecks. *Operations Research*, **60** (2012), No. 3, 716-738.

[20] Ye, H.Q. and D.D. Yao, Diffusion Limit of Fair Resource Control — Stationary and Interchange of Limits, *Mathematics of Operations Research*, **41** (2016), No. 4, 1161-1207.

[21] Ye, H.Q. and D.D. Yao, Justifying Diffusion Approximations for Stochastic Processing Networks under a Moment Condition. *Annals of Applied Probability*, **28** (2018), 3652-3697.