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
© 0000 INFORMS

Diffusion Limit of Fair Resource Control — Stationarity and Interchange of Limits

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

We study a resource-sharing network where each job requires the concurrent occupancy of a subset of links (servers/resources), and each link's capacity is shared among job classes that require its service. The real-time allocation of the service capacity among job classes is determined by the so-called "proportional fair" scheme, which allocates the capacity among job classes taking into account both the queue lengths and the shadow prices of link capacity. We show that the usual traffic condition is necessary and sufficient for the diffusion limit to have a stationary distribution. We also establish the uniform stability of the pre-limit networks; and hence, the existence of their stationary distributions. To justify the interchange of two limits, the limit in time and the limit in diffusion scaling, we identify a *bounded-workload* condition, and show it is a sufficient condition to justify the interchange for both the stationary distributions and their moments. This last result is essential for the validity of the diffusion limit as an approximation to the stationary performance of the original network. We present a set of examples to illustrate justifying the validity of diffusion approximation in resource-sharing networks, and also discuss extensions to other multi-class networks via the well-known Kumar-Seidman/Rybko-Stolyar model.

Keywords: stochastic processing network, proportional fair allocation, diffusion limit, stationary distribution, interchange of limits, uniform stability.

1. Introduction In classical queueing control, the optimal policy often takes the form of a static or dynamic priority rule. The primary examples are the well-known $c\mu$ rule (static); and the "Gittins index" rule (dynamic). The limitation of these rules is two-fold. First, the priority rule, which commits at any time the entirety (100%) of the resource to the job class with the highest priority, may be neither practical nor desirable in certain applications. For instance, in many service systems, the real-time allocation of resources must observe some notion of "fairness" among the various classes of jobs or customers. Consequently, resources are always *shared*, with suitable weighting schemes to differentiate the classes, but even the lowest ranked jobs will get some allocation. Second, the optimality of these rules are mostly limited to a single server or a stand-alone service facility as opposed to a network of resources. In particular, the models do not allow features such as concurrent occupancy of multiple resources that are distributed and inter-connected in a network, what's known as a *stochastic processing network* (Harrison [29, 30]).

We study a stochastic processing network called resource-sharing network (e.g., [31]), in which each job requires the concurrent occupancy of a subset of links (servers/resources) that depends on the class identify of the job, whereas each link's capacity is shared at any time by jobs from various classes that require its service. The real-time allocation of the service capacity among job classes is determined by a control scheme, or *protocol*, the so-called "proportional fair" allocation scheme, which allocates the service capacity among job classes taking into account both the congestion levels (queue lengths) and the shadow prices for consuming link capacities. This has been a popular model to study congestion control on the Internet (refer to Bonald and Massoulie [3], de Veciana *et al* [18], Kelly *et al* [37], Massoulie and Roberts [43], Mo and Walrand [45], and many others). Mathematically, proportional fair allocation belongs to the class of so-called utility-maximizing control — it maximizes a log-utility objective that is a function of the network state.

This type of dynamic resource control, however, makes it very challenging to evaluate the performance of the network. Even in the setting of Poisson arrivals and exponential service times, in which the queue-length process can be modeled as a continuous-time Markov chain, the transition rate from one state to another is itself a solution to an optimization problem, making the Markov chain a rather intractable object to analyze. Therefore, it is quite impossible to tell (unless one resorts to simulation) what the resource control scheme achieves in terms of the system performance over any extended period of time, although we do know in each state the protocol maximizes a given fairness measure.

1.1. Background Research in recent years has demonstrated the effectiveness of applying fluid and diffusion scalings to the network and investigating the corresponding limiting regime. Specifically, the stability of resource-sharing network under the proportional fair allocation was established via fluid models; refer to Bonald and Massoulie [3], Kelly and Williams [38] and Ye *et al* [57], among others; in particular, the notion of an *invariant* (or *fixed-point*) state was developed in [38] — if the fluid model starts in such a state it will always stay in that state. Kang *et al* [32] have established the diffusion limit of the network with multiple bottlenecks (a server is a bottleneck if its capacity equals the total nominal workload of all job classes that require its service), under the proportional fair control scheme, but requiring the additional condition that every server in the network has a dedicated local traffic — a job class that uses only that server and no other servers. In [60], we have established the diffusion limit for the same network and without the local traffic condition assumed in [32]. What we require is a substantially weaker condition, that the constituent matrix, which maps the link occupancies (rows) to routes (columns), be full row-rank.

In both [32] and [60], the diffusion limit is characterized by the so-called dynamic complementarity problem (DCP), also known as Skorohod problem. With further conditions on the network parameters, namely the reflection and covariance matrices satisfying the so-called skew-symmetry condition, the diffusion limit will have a stationary distribution with a product-form density function. This has motivated us to study the stationary distribution of the diffusion limit in the general setting. For instance, under what conditions the stationary distribution exists; and how does the stationary distribution relate to the stationary distributions of the original network and its diffusion-scaled, pre-limit versions.

These questions are succinctly and precisely captured in Figure 1 (also refer to similar figures in [23, 26]). Consider a sequence of networks under heavy traffic, with each network involving the proportional fair allocation of concurrent resources as described above. The process in question is the workload at time t, $\hat{W}^k(t)$, a vector process in the k-th diffusion-scaled network. As $k \to \infty$, the diffusion limit $\hat{W}(t)$ as established in [32] and [60] is represented by the left vertical side, the side I, of the rectangle. The goal of this paper is to establish the other three sides of the rectangle.

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

Figure 1. Interchange of limits

First, we will prove the existence of the stationary distribution (i.e., as $t \to \infty$) for both the diffusion limit $\hat{W}(t)$ and the pre-limit $\hat{W}^k(t)$; and we do so under the *usual traffic condition* of the diffusion limit (plus other standard conditions required for the diffusion limit). These correspond to the two horizontal sides, II and III, of the rectangle. Next, we will show that the stationary distribution of the pre-limit process, under diffusion scaling (i.e., as $k \to \infty$), converges weakly to the stationary distribution of the diffusion limit, under a *bounded workload condition.* This is represented by side IV of the rectangle. Notably, this last result affirms that under the bounded workload condition letting $t \to \infty$ and $k \to \infty$, in either order, will lead to the same limit, and hence validates the so-called interchange of limits. This is essential to justifying the diffusion limit as a valid approximation for the stationary performance of the original or pre-limit networks.

This type of justification has been established for the generalized Jackson network by Gamarnik and Zeevi [23], and by Budhiraja and Lee [6]. (More precisely, in this setting only side IV was a new result, the other three sides had been previously proven.) The generalized Jackson network, being a single-class model, has several notable advantages. The primary one is that the Skorohod problem, which characterizes the dynamics of the pre-limit networks, is essentially the same — modulo the scaling constants — as the one that governs the diffusion limit; in particular, the complementarity (or, work conserving) condition holds in both. In addition, the reflection matrix is an M-matrix, which in the single-class setting results in the Lipschitz continuity of the Skorohod mapping. Consequently, the interchange of limits can be established under the usual traffic condition, along with the standard technical conditions on the moments of the interarrival and service times — conditions that are also required for the diffusion limit.

Other recent studies that justify this type of interchange of limits include: Katsuda [34, 35] for a multi-class single-server queue, Tezcan [52] for a multi-server pool with a single job class, Gamarnik and Goldberg [21] for the M/M/N queue, Gamarnik and Stolyar [22] for a multi-server pool with multiple job classes, and Dai *et al* [13] for the many-server queues with abandonment. The last four papers involve the Halfin-Whitt regime. Earlier related works include the following. Kaspi and Mandelbaum [33] showed that in an irreducible closed network the scaled stationary distribution converges to the stationary distribution of a reflected Brownian motion on a simplex;

the boundedness of the scaled queue lengths is the distinct feature that leads to the tightness of the stationary distributions of pre-limit networks and then the interchange of limits. Harrison [27, 28] considered the steady-state waiting-time distributions in a sequence of tandem queueing systems under heavy traffic, and showed that under diffusion scaling, the sequence of stationary distributions converges weakly to the stationary distribution of the diffusion limit. His approach made use of the explicit expression of the waiting time in terms of interarrival times and service times. Szczotka and Kelly [51] established the same result, allowing certain dependency among the service times. Their approach was based on a representation of the (steady-state) waiting time in terms of a two-sided stationary extension of the sequence of interarrival and service times. Notably, the Lindley-type of recursion that connects the waiting time to the interarrival and services times, which is the key to the approaches in both works, is *not* present in networks with a more complex configuration or with multiple job classes.

1.2. Overview of This Study In a multi-class network, such as the one we focus on here, the main difficulty often lies in the fact that *complementarity*, in general, does not hold for the pre-limit network. In fact, it only holds in an "approximate" sense as follows: a link (server) will not be idle unless the network state (represented by queue lengths or workloads) approaches the facet of the fixed-point state space associated with the link. (The fixed-point state space is the area towards which the utility-maximizing allocation will drive the state process of the network, under diffusion scaling; and this is where the diffusion limit resides.) Indeed, this is the key property one needs to establish so as to prove the diffusion limit.

To show that the diffusion limit has a stationary distribution (side III), we prove that the usual traffic condition is sufficient (and necessary) for the fluid model $\hat{w}(t)$, the deterministic counterpart of the diffusion limit $\hat{W}(t)$ (i.e., replacing the free process in $\hat{W}(t)$ by its drift), to be stable; refer to Theorem 8(a). The stability of $\hat{w}(t)$ is then connected to the stationary distribution of the diffusion limit via the approach of Dupuis and Williams [19]; refer to Theorem 4. These results are depicted in the bottom part of Figure 2, where $A\theta < 0$ is the usual traffic condition, and the arrows denote implication relations.

$$\begin{tikzcd}
\hat{W}^k(t)\text{: stable} & \hat{w}^k(t)\text{: uniformly stable} \arrow[l, Rightarrow] & & \\
\hat{W}(t)\text{: stable} & \hat{w}(t)\text{: stable} \arrow[l, Rightarrow] \arrow[u, Rightarrow] & \Big(\arrow[r, Leftrightarrow] & A\theta < 0 \ \Big)
\end{tikzcd}$$

Figure 2. Relations among stability results

In fact, the stability of $\hat{w}(t)$ also plays a key role in side II of Figure 1, the existence of stationary distributions of the pre-limit networks. It turns out that the key is the *uniform stability* of the pre-limit networks, or their deterministic counterparts $\hat{w}^k(t)$ (Theorem 7). Roughly, this means, for sufficiently large k (the scaling constant) the workload $\hat{w}^k(t)$ associated with any such network

k, starting with a total initial workload that is bounded (by a single unit, say), can be drained by a time that is independent of k. We show this strengthened form of stability, too, holds under the stability of $\hat{w}(t)$. Refer to the vertical arrow in Figure 2. The key to such a result is the convergence, $\hat{w}^k(t) \to \hat{w}(t)$, the proof of which essentially follows the proof of the diffusion limit (side I in Figure 1), with the removal of randomness. The relations in Figure 2 have brought out the central role played by the stability of $\hat{w}(t)$.

A recent study by Gurvich [26] has made clear that in multi-class networks, the interchange of limits, side IV in Figure 1, will require additional conditions (i.e., above and beyond the conditions that lead to the stationary distributions corresponding to sides II and III). The approach in [26] is to relate the interchange of limits to a fluid model, and present a sufficient condition for the interchange as requiring any solution to this fluid model to converge to the fixed-point state space at a linear rate. Several remarks are in order. First, while it is clear that some condition is needed to justify the interchange, characterizing the condition is highly non-trivial, and the difficulties involved have been convincingly illustrated in [26] (as well as in our study here). Second, Gurvich considers a class of networks under the queue-ratio service discipline, which is very different from the type of simultaneous resource-occupancy network under proportional fair allocation that is the main subject of our study here. (In §6, however, we show that our bounded workload condition can also be applied to the well-known Kumar-Seidman network, which operates under a priority service discipline.) Third, our bounded workload condition can be viewed as a relaxed version of Lipschitz continuity, the latter is the key to justifying the interchange of limits in the single-class generalized Jackson network (refer to Theorem 3.3 of Budhiraja and Lee [6]), but usually will not hold in multi-class networks. Refer to more details at the beginning of §4.

Another recent study by Shah *et al* [48] proves the interchange of limits for the same network model as ours, but with the additional restriction of Poisson arrivals and i.i.d. exponential service times. Making use of the distributional (exponential) information, the authors are able to first bound the one-step transition of a Lyapunov function (which is basically the mean drift of a Foster criterion), and then connect to its running maximum via Doob's maximal inequality, leading to a uniform bound for all prelimit networks. In contrast, without the Markovian advantage associated with the exponential (interarrival and service) times, we must deal with the sample paths directly (as did Budhiraja and Lee [6], and earlier, Dai and Meyn [15]). In this type of sample-path approach, requiring certain condition to ensure the uniform integrability appears to be unavoidable — refer to similar cases in, e.g., Dai ([12], Lemma 4.5) and Dai and Meyn ([15], Lemma 5.2). Yet, in these studies, the arrival processes themselves can serve as bounds for the workloads and directly yield the uniform integrability; whereas in our case the diffusion-scaled arrival processes are unbounded, thus requiring extra condition (such as our bounded workload condition).

1.3. Main Results and Contributions

For ease of reading, all results (lemmas, propositions and theorems) are numbered consecutively in the order they appear. Four theorems and two propositions constitute the main results in the paper, which are highlighted below:

- Theorem 4 in §3.1 establishes that the stability of $\hat{w}(t)$ is sufficient for the diffusion limit to have a stationary distribution. Theorem 7 in §3.2 establishes the uniform stability of fluid models associated with the pre-limit networks, and hence, their stationary distributions and moments of the pre-limit networks, also under the stability of $\hat{w}(t)$.

- Theorem 8 in §3.3 refines the above results for the resource-sharing network, and shows that the usual traffic condition is necessary and sufficient for the diffusion limit to have a stationary distribution, and sufficient for the pre-limit networks to have stationary distributions and moments.
- Proposition 11 and Theorem 14 in §4.1 justify the interchange of limits, for both stationary distributions and moments.
- Proposition 15 in §6 verifies the bounded workload condition for the well-known Kumar-Seidman (or, Rybko-Stolyar) network, and thus justifies the interchange of limits for that model.

Our study provides a systematic approach to justify the heavy traffic stationary distribution as a valid approximation for a class of resource-sharing networks under proportional fair control, and possibly for a broader range of other multi-class stochastic processing networks as well. Our approach consists of three steps. First, identify conditions for the stability of the deterministic version of the DCP associated with the diffusion limit, and apply the technique of Dupuis and Williams [19] to claim the existence of the stationary distribution of the diffusion limit. Second, establish the uniform stability of the fluid model corresponding to the pre-limit networks, and this implies the stability of the pre-limit networks (via essentially the same proof for the diffusion limit). Third, verify the bounded workload condition and claim the justification for the interchange of limits.

1.4. Outline of the Paper

The rest of the paper is organized as follows. We start in §2 with details of the resource-sharing network model, followed by a summary of the diffusion limit (established in [60]), along with extensions that modify the condition on the initial states and allow a mixed scaling, both are needed in later proofs. In §3, we study the stationary distributions of both the diffusion limit and the pre-limit networks, and establish the fact that the stability of $\hat{w}(t)$ (or the usual traffic condition) is sufficient for both. For the pre-limit networks, as alluded to above, the key is the uniform stability. In §4, we start with introducing the bounded workload condition, followed by proving the tightness of the workload associated with the pre-limit networks, and justifying the interchange of limits for both stationary distributions and moments. Three examples are presented in §5, where we illustrate how the bounded workload condition can be verified. In addition, a sufficient condition for bounded workload is also discussed to shed more insight to its role. In §6, we demonstrate that our approach has the potential to extend to other multi-class networks. In particular, we verify the bounded workload condition for the well-known Kumar-Seidman network, which, operating under a priority service discipline, lies outside of the resource-sharing network model. Concluding remarks are summarized in §7.

To facilitate the flow of exposition, we take a modularized approach: there is a separate appendix for each of the main sections, where we collect longer proofs as well as secondary technical results. Specifically, each of the appendices, B$\sim$D, serves one of the sections, §2$\sim$§4, *exclusively*, i.e., results in one appendix are not used in any other ones. Additional preliminary or technical results that supplement those in §2 and are used in subsequent sections or cross-referenced in other appendices are collected in Appendix A.

Fluid models	Defining equations	Limits of
$\hat{w}^k(t)$	(56-59)	$\hat{W}^k(mt)/m$ as $m \to \infty$ (Lemma 5)
$\hat{w}(t)$	(39-44)	$\hat{W}^k(m_k t)/m_k$ as $k \to \infty$ (Proposition 3) $\hat{w}^k(t)$ as $k \to \infty$ (Lemma 6)
$\bar{w}(t)$	(130-132)	$\bar{W}^k(t)$ as $k \to \infty$ (Proposition 16) $\bar{W}^{k,j_k}(t)$ as $k \to \infty$ (Lemma 24) $\bar{w}^{k,j_k}(t)$ as $k \to \infty$ (Lemma 25)

Table 1. Fluid models and associated equations

For easy reference, in Figure 3 we summarize the relations among the theorems, propositions and lemmas that lead to the main result, the interchange of limits in Theorem 14. (Secondary technical results are not included in Figure 3.) Moreover, in Table 1, we list the three fluid models used repeatedly in our approach, along with their defining equations. Simply put, the fluid models, $\hat{w}^k(t)$ and $\hat{w}(t)$, are deterministic counterparts of the diffusion-scaled workload process $\hat{W}^k(t)$ and its limit $\hat{W}(t)$; and the third fluid process, $\bar{w}(t)$, is used mainly in the technical proofs. The various ways to reach these fluid models (as limits of scaled workload processes) are summarized in the last column of the table.

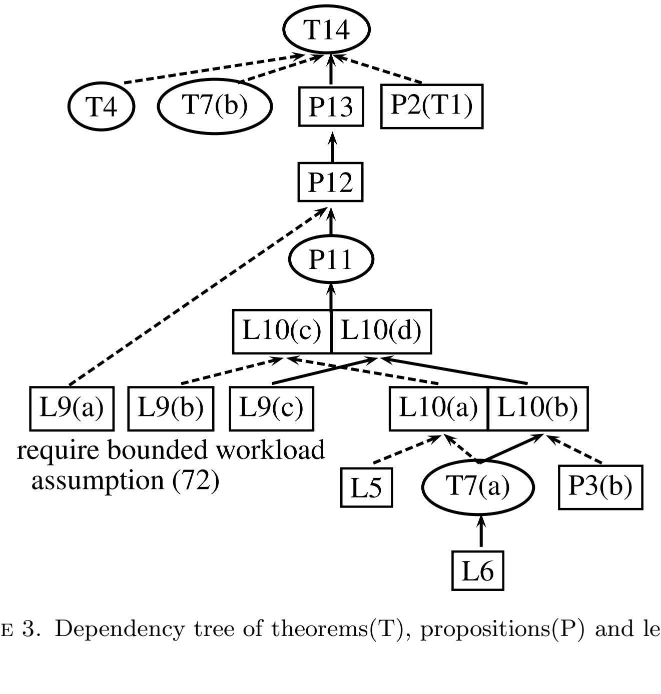

Figure 3. Dependency tree of theorems(T), propositions(P) and lemmas(L)

2. The Resource-Sharing Network, Preliminaries and Extensions The network model under study (as in [60]) consists of a set of servers (or "links") $\mathcal{L}$; and a set of job classes, $\mathcal{R}$, with each class corresponding to a "route" – a subset of links. Denote $\ell \in r$ if link ℓ is part of route r. To be processed in the network, each class r job requires the *simultaneous* occupancy of all the links involved in the route. For the rest of the paper, we shall use the terms "link" and "server" interchangeably, and likewise for "route" and "class".

Denote $R = |\mathcal{R}|$ and $L = |\mathcal{L}|$; and let $A = [a_{\ell r}]_{\ell \in \mathcal{L}, r \in \mathcal{R}}$ be a *non-negative* matrix of dimension $L \times R$. Assume A has a full row-rank of L; hence, $L \leq R$. 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 $a_{\ell r} = \mathbf{1}\{\ell \in r\}$. That is, each row-$\ell$ of A identifies (with an entry 1) all the routes that link ℓ is part of, whereas each column-r of A lists all the links that route r will traverse. The general case of allowing $a_{\ell r}$ to take on non-negative real values extends the scope of the model, for instance, to the setting of multi-path routing, where each route is a source-destination pair that can be connected via multiple paths, i.e., multiple subsets of links; refer to [32, 36].

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 second moments. 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 traffic load (or, traffic intensity) as $\rho = (\rho_r)_{r \in \mathcal{R}}$, with

$$\rho_r = \lambda_r \nu_r. \tag{1}$$

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). In some applications, the network model allows *full* processor-sharing – the server capacities are shared among *all* jobs present in the network (e.g., [46]). In the case of Poisson arrivals and exponential service times, these two models are equivalent – the state processes evolve following the same probabilistic law. Otherwise, our results below do not extend to the full processor-sharing case, which requires keeping track of the residual interarrival and service times, and hence different approaches are needed, such as those associated with measure-valued processes (e.g., Gromoll and Williams [24, 25]).

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 by $\Lambda(n) = (\Lambda_r(n))_{r \in \mathcal{R}}$, 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:

$$\Gamma = \{\gamma = (\gamma_r)_{r \in \mathcal{R}} : \; A\gamma \leq c, \; \gamma \geq 0\}. \tag{2}$$

We assume the *proportional fair allocation* is followed, i.e., given the weights $\beta_r > 0$ ($r \in \mathcal{R}$), $\Lambda(n)$ is the solution to the following optimization problem:

$$\max_{\gamma \in \Gamma} \sum_{r \in \mathcal{R}} \beta_r n_r \log(\gamma_r). \tag{3}$$

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

$$E_r(t) = \max\left\{ i : \sum_{j=1}^{i} u_r(j) \leq t \right\} \quad \text{and} \quad S_r(t) = \max\left\{ i : \sum_{j=1}^{i} v_r(j) \leq t \right\}. \tag{4}$$

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

$$E^0_r(t) = \max\left\{ i-1 : \sum_{j=2}^{i} u_r(j) \leq t \right\}, \quad \text{and} \quad S^0_r(t) = \max\left\{ i-1 : \sum_{j=2}^{i} v_r(j) \leq t \right\}. \tag{5}$$

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

$$N_r(t) = N_r(0) + E_r(t) - S_r(D_r(t)), \tag{6}$$

$$D_r(t) = \int_0^t \Lambda_r(N(s))ds. \tag{7}$$

Note that $S_r(D_r(t))$ represents the total number of class-r service completions by time t.

For much of the analysis below, we find it more convenient to focus on the nominal workload (or workload for short), rather than the queue length, associated with each route:

$$W_r(t) = \nu_r N_r(t), \qquad t \geq 0,\ r \in \mathcal{R}. \tag{8}$$

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. Note $W_r(t)$ as defined above is more precisely the amount of capacity required to serve all class-r jobs that are present in the system at time t, and thus a proxy of the actual workload (cf. Harrison [31]),

We follow the standard approach (e.g., [6, 12, 23, 26]) 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:

$$U_r(t) = \sum_{i=1}^{E_r(t)+1} u_r(i) - t, \qquad V_r(t) = \sum_{i=1}^{S_r(D_r(t))+1} v_r(i) - D_r(t). \tag{9}$$

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 real space, denoted by $\mathcal{X}$ (cf. [12, 17, 34]). 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 $W(t)$, we will write $\Xi(t;x)$ and $W(t;x)$, with $x=\Xi(0)\in\mathcal{X}$ being the initial state.

The above Markov representation of the network is necessary for much of the proofs below, which rely heavily on the theory of Markov processes. It would be useful, however, to keep in mind that in the special case of Poisson arrivals and exponential service times, the workload $W(t)$ per se is already a Markov process, instead of the more elaborate $\Xi(t)$. Focusing on this special case, as the reader may choose to do below, has the advantage of getting directly to the main ideas, without the interference of all the technicalities involving the appended states $U(t)$ and $V(t)$.

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 their indices k omitted); but they may differ in their arrival rates and mean service times, which are also indexed by k (such as λ_r^k, ν_r^k, $\sigma_{a,r}^k$, $\sigma_{s,r}^k$, and $\rho_r^k=\lambda_r^k\nu_r^k$). We assume the existence of the following limits of key parameters, as $k\to\infty$:

$$(\lambda_r^k,\nu_r^k,\sigma_{a,r}^k,\sigma_{s,r}^k)\to(\lambda_r,\nu_r,\sigma_{a,r},\sigma_{s,r}) \quad\text{and}\quad k(\rho_r^k-\rho_r)=k(\lambda_r^k\nu_r^k-\lambda_r\nu_r)\to\theta_r, \qquad r\in\mathcal{R}. \tag{10}$$

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 link ℓ is called a bottleneck (link), if $A_\ell\rho=\sum_{r\in\mathcal{R}}a_{\ell r}\rho_r=c_\ell$, i.e., the total traffic load on that link is equal to its capacity (asymptotically). Below, for ease of exposition, we shall assume that *all* links in the network are bottleneck links, and hence, the following heavy traffic condition:

$$A_\ell\rho=c_\ell, \qquad \ell\in\mathcal{L}. \tag{11}$$

This all-bottleneck condition will allow us to avoid excessive notation in keeping separate accounts for the non-bottleneck links and non-bottleneck routes; the latter involve non-bottleneck links only and typically have zero workload in the limiting regime. On the other hand, all results below extend readily to networks with both bottleneck links and non-bottleneck links, following the steps similar to those outlined in [60].

Recall, we require the primitives of the network, the interarrival times and service requirements, to possess a finite second moment. Now we have a *sequence* of networks, we need to strengthen this condition so that it holds *uniformly* for all the networks. To avoid technicality, 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,

$$\lambda_r^k u_r^k(i) = \lambda_r^1 u_r^1(i) \ \text{ and } \ (\nu_r^k)^{-1} v_r^k(i) = (\nu_r^1)^{-1} v_r^1(i), \qquad i \geq 2, \ r \in \mathcal{R}. \tag{12}$$

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

$$\mathsf{E} \sum_{r \in \mathcal{R}} [(u_r^1(2))^p + (v_r^1(2))^p] < \infty. \tag{13}$$

(Note that to guarantee the convergence in Theorem 1 below, it suffices to have $p = 2$; requiring $p > 2$ is for justifying the interchange of limits later.)

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

$$\mathsf{P}\{u_r^1(2) \geq a\} > 0 \ \text{ for any } a > 0; \tag{14}$$

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

$$\mathsf{P}\left\{ a \leq \sum_{i=2}^{j} u_r^1(i) \leq b \right\} \geq \int_a^b p(x) dx, \quad \text{for any } 0 \leq a < b. \tag{15}$$

These are technical assumptions required for the ergodicity of pre-limit networks later in Theorem 7. They also appeared in prior works, e.g., [12, 5].

To characterize the diffusion limit below, we follow the same approach in [60] to introduce the fixed-point state space and related matrices. Let $B = \operatorname{diag}(b_r)_{r \in \mathcal{R}}$, 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

$$\mathcal{W} = \{ w : w = BA^T \pi, \ \pi = (\pi_\ell)_{\ell \in \mathcal{L}} \geq 0 \}, \tag{16}$$

which is an L-dimensional polyhedral cone in the positive orthant of the R-dimensional real space. From (3), it is clear that $\mathcal{W}$ contains all states in which $\Lambda_r(n) = \rho_r$ (for $n_r > 0$); and indeed this is the space where the diffusion limit (described below) resides. Define an $R \times (R - L)$ matrix H:

$$ABH = 0 \ \text{ and } \ H^T BH = I. \tag{17}$$

Define $G = A^T (ABA^T)^{-1}$. Then, we have

$$ABG = I \ \text{ and } \ G^T BH = 0. \tag{18}$$

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

$$w = BGy + BHz, \quad \text{with } y = Aw, \ z = H^T w, \tag{19}$$

or alternatively,

$$w = BA^T \pi + BHz, \quad \text{with } \pi = G^T w, \ z = H^T w. \tag{20}$$

Moreover, it can be observed that

$$w \in \mathcal{W} \quad \text{if and only if} \quad G^T w \geq 0,\ H^T w = 0. \tag{21}$$

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

$$d^{fp}(w) = \sum_{\ell \in \mathcal{L}} (-g_\ell^T w)^+ + \sum_{m=1}^{R-L} |h_m^T w|. \tag{22}$$

Clearly, w is a fixed-point 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 (6-9) with the index k suitably appended.

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

$$\begin{aligned}
(\hat{E}_r^{0,k}(t), \hat{S}_r^{0,k}(t)) &= \frac{1}{k}\left(E_r^{0,k}(k^2t) - \lambda_r^k k^2 t, S_r^{0,k}(k^2t) - (\nu_r^k)^{-1} k^2 t\right),\\
(\hat{E}_r^{k}(t), \hat{S}_r^{k}(t)) &= \frac{1}{k}\left(E_r^{k}(k^2t) - \lambda_r^k k^2 t, S_r^{k}(k^2t) - (\nu_r^k)^{-1} k^2 t\right),\\
(\hat{\Xi}_r^k(t), \hat{N}_r^k(t), \hat{W}_r^k(t)) &= \frac{1}{k}\left(\Xi_r^k(k^2t), N_r^k(k^2t), W_r^k(k^2t)\right).
\end{aligned}$$

Then, we can rewrite the dynamics in (6-8) as follows:

$$\begin{aligned}
\hat{W}^k(t) &= \text{diag}(\nu)\hat{N}^k(t) = \hat{W}^k(0) + \hat{X}^k(t) + k[\rho t - \tilde{D}^k(t)] \qquad (23)\\
&= \hat{W}^k(0) + \hat{X}^k(t) + BG\hat{Y}^k(t) + BH\hat{Z}^k(t);\\
\tilde{D}^k(t) &= \int_0^t \Lambda(\hat{N}^k(s))ds; \qquad (24)\\
\hat{X}_r^k(t) &= \nu_r^k\left(\hat{E}_r^k(t) - \hat{S}_r^k(\tilde{D}_r^k(t))\right) + k(\rho_r^k - \rho_r)t, \quad \text{for } r \in \mathcal{R}; \qquad (25)\\
\hat{Y}^k(t) &= kA[\rho t - \tilde{D}^k(t)] = k[ct - A\tilde{D}^k(t)], \text{ is non-decreasing in}\, t \geq 0, \quad \text{for } \ell \in \mathcal{L}; \qquad (26)\\
\hat{Z}^k(t) &= kH^T[\rho t - \tilde{D}^k(t)]. \qquad (27)
\end{aligned}$$

The process, $\tilde{D}_r^k(t) = D_r^k(k^2t)/k^2$, is a fluid-scaled process. Note that to obtain (24), we have used the the radial homogeneity property of $\Lambda(n)$: $\Lambda(\alpha n) = \Lambda(n)$ for any constant $\alpha > 0$; and the second equality in (23) follows from applying the decomposition in (19) to the term $k[\rho t - \tilde{D}^k(t)]$. Denote $\hat{X}^k(t) = (\hat{X}_r^k(t))_{r \in \mathcal{R}}$.

All the processes above, primitive or derived, belong to the $\mathcal{D}$-space, the space of functions that are right continuous with left limits (RCLL). Below, we study the weak convergence (or convergence in distribution, denoted as "⇒") of the diffusion-scaled processes. To this end, strictly speaking, we need to work with the Skorohod metric ([1]). However, since all the limiting processes involved are continuous processes (Brownian motions), it is convenient (and indeed equivalent) to continue treating the $\mathcal{D}$-space as endowed with the more familiar uniform metric and uniform convergence on compact set (u.o.c. convergence).

For the derived processes, denote their limits as follows:

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

The existence of these limits is part of the next theorem. Furthermore, the limiting processes are characterized by the following DCP:

$$\hat{W}(t) = \hat{W}(0) + \hat{X}(t) + BG\hat{Y}(t) + BH\hat{Z}(t) \ (\geq 0), \qquad \text{for } t \geq 0; \tag{28}$$
$$G^T\hat{W}(t) \geq 0, \qquad \text{for } t \geq 0; \tag{29}$$
$$\hat{Y}_\ell(t) \text{ is non-decreasing in } t \geq 0,\ \hat{Y}_\ell(0) = 0, \qquad \ell \in \mathcal{L}; \tag{30}$$
$$\int_0^\infty \hat{W}(t)^T G \, d\hat{Y}(t) = 0; \tag{31}$$
$$H^T\hat{W}(t) = 0, \qquad \text{for } t \geq 0; \tag{32}$$
$$\hat{Z}(0) = 0; \tag{33}$$

where $\hat{W}(0) \in \mathcal{W}$ 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 (10), and covariance (matrix)

$$\Upsilon = \mathrm{diag}(\sigma_r^2)_{r\in\mathcal{R}}, \qquad \text{with} \qquad \sigma_r^2 = \nu_r^2(\lambda_r^3\sigma_{a,r}^2 + \rho_r\nu_r^{-3}\sigma_{s,r}^2) = \lambda_r\nu_r^2(\lambda_r^2\sigma_{a,r}^2 + \nu_r^{-2}\sigma_{s,r}^2). \tag{34}$$

Theorem 1 (Diffusion Limit [60]) Suppose the heavy-traffic condition in (11) is in force; and under the diffusion scaling, the initial workloads converge to some (random) fixed-point state, while the (time-zero) residuals vanish:

$$\hat{W}^k(0) \Rightarrow \hat{W}(0) \in \mathcal{W}, \tag{35}$$
$$|\hat{U}^k(0)| + |\hat{V}^k(0)| = \frac{1}{k}(|u^k(1)| + |v^k(1)|) \to 0, \quad a.s. \tag{36}$$

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

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

with the limit characterized by the equations in (28-33).

Note the above diffusion limit theorem serves as the starting point of our study of the other three sides in Figure 1. In this regard, prior results leading to this theorem are relevant to this study as well, and these are detailed in Appendix A. Here we highlight several key points:

- Uniform attraction: The fluid model $\bar{w}(t)$ (see Table 1), converges to a fixed point state as $t \to \infty$ (Proposition 18). Intuitively, $\bar{w}(t)$ is the deterministic version of the a critically loaded network, where the offered traffic for each server matches its capacity. The uniform attraction is the key property ensuring state-space collapse, i.e., the fixed-point state space $\mathcal{W}$ is of a lower dimension (L) than the original state space (of dimension R).
- Complementarity and oscillation inequality: As mentioned in the introduction, there is a reflection force at the boundary of the fixed-point state space $\mathcal{W}$ (Lemma 19); consequently, complementarity holds in an "approximate" sense, which is key to establishing the above diffusion limit. Complementarity is also required here to invoke the oscillation inequality (Lemma 21, or originally, Proposition 7.1 of [32]) to establish the boundedness (or tightness) of the pre-limit networks.

To proceed further, i.e., to establish the stationarity (sides II and III in Figure 1) and to validate the interchange of limits (side IV), we need variations of the above diffusion limit theorem, which we present next, along with pointers to exactly where they will be used later.

For the convergence to hold along the *full* sequence of k, it is necessary to assume the initial condition in (35); specifically, the limit $\hat{W}(0)$ must be a fixed-point state so that (29) and (32) are satisfied at $t=0$. As remarked by Bramson [4] (also see [42, 50]), this initial condition can be relaxed (to reach a weaker conclusion). The next proposition illustrates such a variation: instead of assuming the convergence in (35), we assume the sequence of initial states (including the residuals) is tight. The result will be used in the proof of Theorem 14.

Proposition 2 Suppose the heavy-traffic condition in (11) is in force; and the sequence of initial states $\{\hat{\Xi}^k(0)\}$ is tight. Let $\{t_0^k\}$ be any sequence of times such that $t_0^k \to 0$ and $kt_0^k \to \infty$ as $k\to\infty$. Then, for any subsequence of k, there exists a further subsequence, denoted by $\mathcal{K}$, such that the following weak convergence holds when $k\to\infty$ along $\mathcal{K}$:

$$\begin{aligned}&\left(\hat{W}^k(t_0^k+t), \hat{X}^k(t_0^k+t)-\hat{X}^k(t_0^k), \hat{Y}^k(t_0^k+t)-\hat{Y}^k(t_0^k), \hat{Z}^k(t_0^k+t)-\hat{Z}^k(t_0^k)\right)\\ &\Rightarrow \left(\hat{W}(t), \hat{X}(t), \hat{Y}(t), \hat{Z}(t)\right),\end{aligned}$$

where the limit follows the specifications in (28-33). Furthermore, we have

$$\hat{W}^k(t_0^k) \Rightarrow \hat{W}(0) \in \mathcal{W}, \quad \text{as } k\to\infty \text{ along } \mathcal{K}; \tag{37}$$

and, for any $M\geq 0$,

$$\limsup_{k\to\infty, k\in\mathcal{K}} \mathsf{P}\{\kappa_w|\hat{\Xi}^k(0)|\leq M\} \leq \mathsf{P}\{|\hat{W}(0)|\leq M\}; \tag{38}$$

where κ_w is a constant that depends only on network parameters (as specified in Proposition 18(a)).

Another variation of the diffusion limit theorem concerns a modified scaling. Let $\{m_k\}$ be any sequence that increases to infinity, i.e.,

$$m_k \to \infty \quad \text{ as } k\to\infty.$$

We then apply the fluid scaling, i.e., with both time and space further scaled by m_k, to the diffusion-scaled processes:

$$\frac{1}{m_k}\left(\hat{W}^k(m_k t), \hat{X}^k(m_k t), \hat{Y}^k(m_k t), \hat{Z}^k(m_k t)\right).$$

Clearly, the net effect of this new "mixed scaling" is that the (original) processes in question, say, the workload $W^k(t)$, is compressed more (than the square root of the time scale). Interestingly, the limiting regime under this mixed scaling appears to be a mixture of the fluid limit and the diffusion limit. Specifically, it follows the DCP in (28-33), but with the (unreflected) Brownian motion $\hat{X}(t)$ replaced by its drift term θt:

$$\hat{w}(t) = \hat{w}(0)+\theta t + BG\hat{y}(t) + BH\hat{z}(t) \ (\geq 0), \quad \text{ for } t\geq 0; \tag{39}$$

$$G^T\hat{w}(t)\geq 0, \quad \text{ for } t\geq 0; \tag{40}$$

$$\hat{y}_\ell(t) \text{ is non-decreasing in } t\geq 0,\ \hat{y}_\ell(0)=0, \quad \ell\in\mathcal{L}; \tag{41}$$

$$\int_0^\infty \hat{w}(t)^T G\ d\hat{y}(t) = 0; \tag{42}$$

$$H^T\hat{w}(t) = 0, \quad \text{ for } t\geq 0; \tag{43}$$

$$\hat{z}(0) = 0. \tag{44}$$

Note that given the initial state $\hat{w}(0) \in \mathcal{W}$, the above is a *deterministic* DCP. Like its stochastic counterpart, $\hat{w}(t)$ also evolves within the fixed-point state space $\mathcal{W}$ following the observation in (21). The next proposition has two parts, corresponding to Theorem 1 and Proposition 2, respectively, under the mixed scaling. The results will be used in establishing Lemma 10. (Proofs of the two propositions can be found in Appendix B.2 and Appendix B.2.)

Proposition 3 Suppose the heavy-traffic condition in (11) is in force.
(a) Assume the initial state under the mixed-scaling converges to some fixed-point state:

$$\begin{aligned}
&\frac{1}{m_k}\hat{W}^k(0) \Rightarrow \hat{w}(0) \in \mathcal{W},\\
&\frac{1}{m_k}(|\hat{U}^k(0)| + |\hat{V}^k(0)|) = \frac{1}{km_k}(|u^k(1)| + |v^k(1)|) \to 0, \quad a.s.
\end{aligned}$$

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

$$\frac{1}{m_k}\left(\hat{W}^k(m_k t), \hat{X}^k(m_k t), \hat{Y}^k(m_k t), \hat{Z}^k(m_k t)\right) \Rightarrow (\hat{w}(t), \theta t, \hat{y}(t), \hat{z}(t)),$$

where the limit follows the specifications in (39-44).
(b) Assume that the sequence of initial states $\{\hat{\Xi}^k(0)/m_k\}$ is tight. Let $\{t_0^k\}$ be any sequence of times such that $t_0^k \to 0$ and $kt_0^k \to \infty$ as $k \to \infty$. Then, for any subsequence of k, there exists a further subsequence, denoted by $\mathcal{K}$, such that the following weak convergence holds when $k \to \infty$ along $\mathcal{K}$:

$$\begin{aligned}
&\frac{1}{m_k}(\hat{X}^k(m_k(t_0^k + t)) - \hat{X}^k(m_k t_0^k)) \Rightarrow \theta t, \quad \text{and}\\
&\frac{1}{m_k}\left(\hat{W}^k(m_k(t_0^k + t)), \hat{Y}^k(m_k(t_0^k + t)) - \hat{Y}^k(m_k t_0^k), \hat{Z}^k(m_k(t_0^k + t) - \hat{Z}^k(m_k t_0^k))\right)\\
&\quad \Rightarrow (\hat{w}(t), \hat{y}(t), \hat{z}(t)),
\end{aligned}$$

where the limit follows the specifications in (39-44). Furthermore, we have

$$\frac{1}{m_k}\hat{W}^k(m_k t_0^k) \Rightarrow \hat{w}(0) \in \mathcal{W}, \quad \text{as } k \to \infty \text{ along } \mathcal{K}; \tag{45}$$

and, for any $M \geq 0$,

$$\limsup_{k \to \infty, k \in \mathcal{K}} \mathsf{P}\{\kappa_w |\hat{\Xi}^k(0)/m_k| \leq M\} \leq \mathsf{P}\{|\hat{w}(0)| \leq M\}; \tag{46}$$

where κ_w is a constant that depends only on network parameters.

3. Stationary Distributions and Uniform Stability This section is devoted to establishing the results depicted in Figure 2. It is divided into three parts: first, the result in the bottom side (excluding the parentheses) in the figure; second, the top side; and third, the parentheses. Note the first two parts may apply to more general settings as we will discuss later (in §6); whereas the third part, the stability of $\hat{w}(t)$ being equivalent to the usual traffic condition, is specific to the resource-sharing network.

3.1. Stability of the Diffusion Limit To establish the existence and uniqueness of the *stationary* distribution of the diffusion limit $\hat{W}(t)$ in Theorem 1, we follow the approach developed in Dupuis and Williams [19] (Theorem 2.6, in particular). By this approach, the key is to establish the stability of the DCP in (39-44), which is the deterministic version of the DCP in (28-33). We shall refer to the deterministic DCP in (39-44) as *stable*, if there exists a time T such that for any solution with $|\hat{w}(0)| \le 1$, we have $\hat{w}(t) = 0$ for all $t \ge T$.

Theorem 4 If the deterministic DCP in (39-44) is stable, then the diffusion limit $\hat{W}(t)$ in Theorem 1 is positive recurrent and has a unique stationary distribution.

Proof. We first transform the two DCP's mentioned above into the standard format in [19]. Letting $\hat{W}_G(t) = G^T\hat{W}(t)$, along with (18), turns the DCP in (28-33) into:

$$\hat{W}_G(t) = \hat{W}_G(0) + G^T\hat{X}(t) + (G^TBG)\hat{Y}(t), \qquad \text{for } t \ge 0; \tag{47}$$

$$\hat{W}_G(t) \ge 0, \qquad \text{for } t \ge 0; \tag{48}$$

$$\hat{Y}_\ell(t) \text{ is non-decreasing in } t \ge 0,\ \hat{Y}_\ell(0) = 0, \qquad \ell \in \mathcal{L}; \tag{49}$$

$$\int_0^\infty \hat{W}_G^T(t)\ d\hat{Y}(t) = 0. \tag{50}$$

Similarly, letting $\hat{w}_G(t) = G^T\hat{w}(t)$ in (39-44) leads to:

$$\hat{w}_G(t) = \hat{w}_G(0) + G^T\theta t + (G^TBG)\hat{y}(t), \qquad \text{for } t \ge 0; \tag{51}$$

$$\hat{w}_G(t) \ge 0, \qquad \text{for } t \ge 0; \tag{52}$$

$$\hat{y}_\ell(t) \text{ is non-decreasing in } t \ge 0,\ \hat{y}_\ell(0) = 0, \qquad \ell \in \mathcal{L}; \tag{53}$$

$$\int_0^\infty \hat{w}_G^T(t)\ d\hat{y}(t) = 0. \tag{54}$$

Note that the reflection matrix G^TBG in (51) is a P-matrix and hence a completely-S matrix; refer to Kang et al. ([32], Lemma 7.1) and Ye and Yao ([60], Proposition 4). Following Theorem 2.6 of [19], if the solution to this last DCP is *stable* (i.e., there exists a time T such that for any solution with $|\hat{w}_G(0)| \le 1$ such that $\hat{w}_G(t) = 0$ for all $t \ge T$), then, the corresponding diffusion limit $\hat{W}_G(t)$ is positive recurrent and has a unique stationary distribution.

From the condition of the theorem (DCP in (39-44) is stable), we know that the deterministic DCP in (51-54) is indeed stable. Consequently, the process $\hat{W}_G(t)$ is positive recurrent and has a unique stationary distribution; and hence, so does $\hat{W}(t)$, following the equivalent relations:

$$\hat{W}_G(t) = G^T\hat{W}(t) \ \text{ and } \ \hat{W}(t) = BG(ABA^T)\hat{W}_G(t). \tag{55}$$

(Note that the equivalence holds for $\hat{w}$ and $\hat{w}_G$ as well and that $G^TBG = (ABA^T)^{-1}$.) □

Note in the above proof, we have applied a *simplified version* of Theorem 2.6 of [19]. The original version in [19] states (in the context of our model) that the diffusion limit $\hat{W}_G(t)$ is positive recurrent and has a unique stationary distribution if any solution to the DCP in (51-54) is *attracted to the origin*, i.e., there exists a time $T < \infty$ such that for all $t \ge T$, we have $|\hat{w}_G(t)| \le \epsilon$ for arbitrarily small $\epsilon > 0$. Clearly, this attraction to origin is implied by our simpler stability condition (the DCP in (51-54) is stable), and on the other hand, it also implies the latter according to Stolyar [49] and Ye and Chen [56]; in other words, these two versions are equivalent.

3.2. Uniform Stability of Pre-Limit Networks It turns out that the DCP in (39-44) ensures not only the positive recurrence of the diffusion limit $\hat{W}(t)$ (as stated in Theorem 4), but also the positive (Harris) recurrence of the pre-limit networks $\hat{\Xi}^k(t)$. Establishing the latter result is the objective of this subsection. Refer to Figure 2. Our approach is to establish the *uniform* stability of the fluid models $\hat{w}^k(t)$ associated with the pre-limit networks. Specifically, we show in Theorem 7 that there exists a time t_0 (independent of k), such that starting with any initial state $||\hat{w}^k(0)|| \le 1$, we have $\hat{w}^k(t) = 0$ for $t \ge t_0$ and for all k sufficiently large. This is accomplished in several steps detailed below.

First, analogous to the DCP in (39-44), which provides characterization for the stability of $\hat{W}(t)$, Lemma 5 below says that a fluid model corresponding to the k-th (diffusion-scaled) network in (23-27) is characterized by the following equations:

$$\begin{aligned}
\hat{w}^k(t) &= \hat{w}^k(0) - k\cdot\text{diag}(\rho^k)\left(te \wedge \frac{\bar{u}^k(1)}{k}\right) + k\left(\bar{d}^k(t) \wedge \frac{\bar{v}^k(1)}{k}\right) + k\rho^k t - k\bar{d}^k(t)\\
&= \hat{w}^k(0) - k\cdot\text{diag}(\rho^k)\left(te \wedge \frac{\bar{u}^k(1)}{k}\right) + k\left(\bar{d}^k(t) \wedge \frac{\bar{v}^k(1)}{k}\right)\\
&\quad + k(\rho^k - \rho)t + BG\hat{y}^k(t) + BH\hat{z}^k(t); && (56)\\
\bar{d}^k(t) &= \int_0^t \bar{\Lambda}(\hat{n}^k(s))ds; && (57)\\
\hat{y}^k(t) &= k[ct - A\bar{d}^k(t)], \quad \text{is non-decreasing in } t \ge 0; && (58)\\
\hat{z}^k(t) &= kH^T[\rho t - \bar{d}^k(t)]. && (59)
\end{aligned}$$

In the above, e denotes a vector of all unit components; the convention $\hat{w}^k_r(t) \equiv \nu_r \hat{n}^k_r(t)$ applies; and $\bar{\Lambda}(n)$ is defined as

$$\bar{\Lambda}_r(n) = \begin{cases} \Lambda_r(n) & \text{if } n_r > 0,\\ \rho_r & \text{if } n_r = 0. \end{cases} \tag{60}$$

(Note the "hat" and "bar" designations in the above processes, such as $\hat{w}^k(t)$ and $\bar{d}^k(t)$, are in line with the scalings of their stochastic counterparts, such as the diffusion-scaled process $\hat{W}^k(t)$ and the fluid-scaled process $\tilde{D}^k(t)$.)

Lemma 5 Consider the k-th (diffusion-scaled) network as depicted in (23-27), for any fixed k. Let $\{m_i; i = 1, 2, \cdots\}$ be a sequence of numbers such that $m_i \to \infty$ as $i \to \infty$; and let $\{x^i \in \mathcal{X}; i = 1, 2, \cdots\}$ be a sequence of initial states such that $|x^i| \le m_i$ for all i. Then, for any subsequence of positive integers, there exists a further subsequence, denoted by $\mathcal{I}$, such that the following (a.s.) convergence holds as $i \to \infty$ along $\mathcal{I}$,

$$\frac{1}{m_i}\hat{\Xi}^k(0; x^i) = \frac{1}{m_i}\left(\hat{W}^k_r(0), \hat{U}^k_r(0), \hat{V}^k_r(0)\right) \to \left(\hat{w}^k_r(0), \bar{u}^k_r(1), \bar{v}^k_r(1)\right),$$

and

$$\frac{1}{m_i}\left(\hat{W}^k(m_i t), \tilde{D}^k(m_i t), \hat{Y}^k(m_i t), \hat{Z}^k(m_i t)\right) \to \left(\hat{w}^k(t), \bar{d}^k(t), \hat{y}^k(t), \hat{z}^k(t)\right) \qquad \text{u.o.c. of } t \ge 0,$$

where the limit is Lipschitz continuous and is a solution to the fluid model in (56-59), with initial condition

$$|\hat{w}^k(0)| + |\bar{u}^k(1)| + |\bar{v}^k(1)| \le 1.$$

The above lemma is a variation of Proposition 4.2 of [57], the only new feature being the residuals $\bar{u}^k(1)$ and $\bar{v}^k(1)$, which can be dealt with in the same way as in the proof of Proposition 16 in Appendix A (the supplement to §2). Hence, the proof of the lemma is omitted. Note that $\bar{\Lambda}(n)$ is not necessarily in Γ for an arbitrarily given state $n \ge 0$. Yet, for any solution to (56-59), $\bar{\Lambda}(\hat{n}^k(t))$ must belong to Γ at any time t, at which point all processes involved (i.e., $\hat{w}^k(t), \hat{y}^k(t), \hat{z}^k(t), \bar{d}^k(t)$) are differentiable. Indeed, these processes are differentiable almost everywhere since they are Lipschitz continuous.

Second, The next lemma connects the stability of $\hat{w}(t)$ to the uniform stability of $\hat{w}^k(t)$. Note that the u.o.c. convergence in the lemma parallels the weak convergence in Proposition 3(b). The proof of the lemma is in Appendix C (which is similar to the proof of Proposition 3(b), given the previously established properties such as the uniform attraction, complementarity and oscillation inequality.)

Lemma 6 Consider a sequence of $\hat{w}^k(t)$ characterized by (56-59) with $|\hat{w}^k(0)+\bar{u}^k(1)+\bar{v}^k(1)| \le 1$. Let $\{t_0^k\}$ be a sequence of times such that $t_0^k \to 0$, and $kt_0^k \to \infty$ as $k \to \infty$. Then, for any subsequence of k, there exists a further subsequence $\mathcal{K}$ such that the following holds, as $k \to \infty$ along $\mathcal{K}$,

$$\hat{w}^k(t_0^k + t) \to \hat{w}(t) \qquad \text{u.o.c. of } t \ge 0,$$

where $\hat{w}(t)$ is a solution to the DCP in (39-44) with $|\hat{w}(0)| \le \kappa_w$, where κ_w is a constant that depends only on network parameters (as specified in Proposition 18(a)).

Finally, given Lemmas 5 and 6, we establish the uniform stability of the fluid models $\hat{w}^k(t)$ and the stationarity of the pre-limit networks $\hat{W}^k(t)$ in the following theorem.

Theorem 7 Consider the sequence of networks in Theorem 1, and assume that the DCP in (39-44) is stable.
(a) (*Uniform Stability*) There exists a time $t_0 > 0$ such that for any sufficiently large k, any solution $\hat{w}^k(t)$ to the fluid model in (56-59) with $|\hat{w}^k(0)| \le 1$ satisfies the following,

$$\hat{w}^k(t) = 0, \qquad t \ge t_0. \tag{61}$$

(b) Consequently, for any sufficiently large k, $\hat{\Xi}^k(t)$ is positive recurrent and has a unique stationary distribution, denoted by $\hat{\pi}^k$; and the stationary workload has a finite $(p-1)$-th moment, i.e.,

$$\mathsf{E}_{\hat{\pi}^k} |\hat{W}^k(0)|^{p-1} < \infty. \tag{62}$$

Proof. Let $\mathcal{K}$ be any subsequence of k such that

$$\hat{w}^k(t_0^k + t) \to \hat{w}(t) \qquad \text{u.o.c. of } t \ge 0,$$

where the sequence of times $\{t_0^k\}$ and the limit $\hat{w}(t)$ follow the specifications in Lemma 6. As the DCP in (39-44) is stable, for its solution $\hat{w}(t)$ there exists a time t_0' such that $\hat{w}(t) = 0$ for $t \ge t_0'$. Hence, the above u.o.c. convergence implies

$$\hat{w}^k(t_0^k + t) \to 0 \qquad \text{u.o.c. of } t \ge t_0'; \tag{63}$$

with the convergence holding for the *full* sequence of k since the choice of the subsequence $\mathcal{K}$ is arbitrary.

Pick any $\delta > 0$. We claim that there exists an index k_0 such that, for any $k \geq k_0$ and for any solution $\hat{w}^k(t)$ to (56-59) with $|\hat{w}^k(0)| \leq 1$, the following inequality holds,

$$|\hat{w}^k(t_0' + \delta)| \leq \frac{1}{2}. \tag{64}$$

Otherwise, we can find a subsequence $\mathcal{K}'$ of k and some solutions $\hat{w}^k(t)$ with $|\hat{w}^k(0)| \leq 1$, such that for $k \in \mathcal{K}'$, we have

$$|\hat{w}^k(t_0' + \delta)| > \frac{1}{2},$$

which contradicts (63).

Having established the inequality in (64), it is then routine to prove the conclusion in (61), with $t_0 = 2(t_0' + \delta)$, following the proof of Theorem 6.1 in [49] (or the proof of Theorem 2.3 in [56]).

The stability of $\hat{w}^k(t)$ $(k \geq k_0)$ just established, along with Lemma 5, then directly implies the positive recurrence of the Markov process $\hat{\Xi}^k(t)$, as well as the existence and uniqueness of its stationary distribution. The finiteness of the $(p-1)$-th moment of the stationary workload in (62) follows from Theorem 4.1(ii) of Dai and Meyn [15] Note that although the results in [15] are presented in the context of traditional multi-class queueing networks, they extend to more general stochastic processing networks (including our model here), where jobs within each class receive services in the first-come-first-served order and the number of jobs in service is bounded at all times. □

The stability property in Theorem 7(a) is said to be uniform because a single time t_0 applies uniformly to all sufficiently large k. It does imply the second part of the above theorem, namely the positive recurrence of $\hat{\Xi}^k(t)$, and the finite moment of the stationary workload. More importantly, it will be used in the next section (specifically, Lemma 10) when we justify the interchange of limits.

3.3. Resource-Sharing Network under the Usual Traffic Condition We now turn to the part in parentheses in Figure 2, namely, the results in Theorem 7 can be connected directly to the usual traffic condition for resource-sharing networks.

First, observe that the usual traffic condition here takes the form $A\theta < 0$. To see this, from $A\rho = c$, the heavy traffic condition in (11), along with (10), we have

$$k(A\rho^k - c) = A[k(\rho^k - \rho)] \to A\theta. \tag{65}$$

Thus, $A\theta < 0$ implies, for sufficiently large k,

$$A\rho^k = \left(\sum_{r \in \mathcal{R}} a_{\ell r} \rho_r^k \right)_{\ell \in \mathcal{L}} < c, \tag{66}$$

which is, of course, the usual traffic condition. Note, however, $A\theta < 0$ is slightly stronger than the above condition; specifically, it is not implied by the latter unless the gap from c is no smaller than order $1/k$. For example, if $A\rho^k = c - 1/k^2$, then, $A\theta = \lim_{k\to\infty} k(A\rho^k - c) = 0$. Ignoring this minor gap, we shall simply refer to $A\theta < 0$ as the usual traffic condition throughout below.

Theorem 8 (a) The usual traffic condition

$$A\theta < 0, \tag{67}$$

is necessary and sufficient for the deterministic DCP in (39-44) to be stable.
(b) $A\theta < 0$ is a necessary and sufficient condition for the diffusion limit $\hat{W}(t)$ in Theorem 1 to be positive recurrent and to possess a unique stationary distribution.
(c) $A\theta < 0$ is a sufficient condition for the conclusion in (b) of Theorem 7, i.e., for any sufficiently large k, $\hat{\Xi}^k(t)$ is positive recurrent and has a unique stationary distribution, and the stationary workload has a finite $(p-1)$-th moment.

Proof. To establish the stability of the DCP in (39-44) under the condition $A\theta < 0$, as stated in part (a) of the theorem, we shall make use of the following so-called "S-condition" (due to Stolyar [49]): given $|\hat{w}(0)| = 1$,

$$\inf_{t \geq 0} |\hat{w}(t)| < 1, \tag{68}$$

which is shown in [49], Theorem 6.1, to be equivalent to the stability of the DCP governing $\hat{w}(t)$.

First, suppose $A\theta < 0$ holds. Note that by applying the oscillation inequality in Lemma 21 (in the Appendix), any solution to the DCP in (51-54) is Lipschitz continuous. Hence, the solution is differentiable almost everywhere for $t \geq 0$. Let $f(t) = \hat{w}(t)^T B^{-1} \hat{w}(t)/2$. Then, taking derivative with respect to t, we have

$$\dot{f}(t) = \hat{w}(t)^T B^{-1} \dot{\hat{w}}(t) = \hat{w}(t)^T B^{-1} \left[\theta + BG\dot{\hat{y}}(t) + BH\dot{\hat{z}}(t) \right] = \hat{w}(t)^T B^{-1} \theta,$$

where the second equality follows from (39), and the last equality from (42) and (43).

We claim that if $|\hat{w}(0)| = 1$, then the equivalent condition for stability in (68) holds. Suppose to the contrary, we have $|\hat{w}(t)| \geq 1$ for all $t \geq 0$. Consider any regular time $t \geq 0$. Since $\hat{w}(t) \in \mathcal{W}$ according to (21), we can write $\hat{w}(t) = BA^T \pi$ for some $\pi = (\pi_\ell)_{\ell \in \mathcal{L}} \geq 0$. Hence, $|\hat{w}(t)| = |BA^T \pi| \geq 1$, which implies

$$|\pi| \geq \kappa_1 > 0 \tag{69}$$

where κ_1 depends on B and A only. This, along with the condition in (67), leads to

$$\dot{f}(t) = (\pi^T AB) B^{-1} \theta = \pi^T A\theta \leq \max_{\ell'} (A_{\ell'} \theta) \sum_{\ell \in \mathcal{L}} \pi_\ell \leq \max_{\ell'} (A_{\ell'} \theta) \kappa_1 < 0. \tag{70}$$

Moreover, since $|w(0)| = 1$, we have

$$f(0) = \frac{1}{2} w(0)^T B^{-1} w(0) \leq \frac{\max_r b_r}{2}. \tag{71}$$

Putting together (71) and (70), we must have $f(t) < 0$ for some $t > 0$, which contradicts the quadratic form of $f(t)$.

Therefore, the DCP governing $\hat{w}(t)$ is stable, which implies that $\hat{W}(t)$ is positive recurrent and has a unique stationary distribution, following Theorem 4.

Conversely, suppose $A\theta < 0$ does not hold, i.e., $A_{\ell'}\theta \geq 0$ for some $\ell' \in \mathcal{L}$. Then,

$$A_{\ell'}\hat{w}(t) = A_{\ell'}\hat{w}(0) + A_{\ell'}\theta t + \hat{y}_{\ell'}(t) \geq A_{\ell'}\hat{w}(0).$$

Therefore, the DCP cannot be stable. Moreover, for the diffusion limit $\hat{W}(t)$, we have from (28),

$$A_{\ell'}\hat{W}(t) = A_{\ell'}\hat{W}(0) + A_{\ell'}\hat{X}(t) + \hat{Y}_{\ell'}(t).$$

According to the minimality on the (one-dimensional) reflected Brownian motion (RBM) (e.g., Chapter 6 of [8]), the RHS above is bounded from below by an RBM driven by the free process $A_{\ell'}\hat{X}(t)$. (Note that $A_{\ell'}\hat{W}(t)$ and $\hat{Y}_{\ell'}(t)$ need not satisfy the complementarity condition.) Since the lower-bounding RBM has a nonnegative drift ($A_{\ell'}\theta \geq 0$), it cannot be positive recurrent, and thus neither is the diffusion limit $\hat{W}(t)$.

Given the conclusion in (a), the conclusion in (c) follows from Theorem 7(b) immediately. (Alternatively, *specific to the resource-sharing network*, we can invoke the results in [15, 57] to claim the conclusion in (c) under the usual traffic condition in (66) directly.) □

4. Interchange of Limits

The uniform stability in Theorem 7(a) turns out to be a key step in establishing our next main result, side IV in Figure 1. We will first introduce a bounded workload condition, which guarantees the bounded p-th moment of the workload processes. Along with the uniform stability property, the bounded p-th moment condition leads to the uniform p-moment stability of workloads. These properties then lead to the tightness of the stationary distributions and finally, the interchange of limits.

First, the key condition to justify the interchange of limits:

Bounded Workload Condition. There is a constant $\kappa > 0$ such that for any index k and any time $t \geq 0$ (and any sample-path), the following condition holds,

$$\sup_{0\leq s\leq t} |\hat{W}^k(s)| \leq \kappa \left(|\hat{W}^k(0)| + \sup_{0\leq s\leq t} |\hat{X}^k(s)| \right). \tag{72}$$

To motivate the above condition, let's compare it against the Lipschitz continuity (of the Skorohod mapping) in the single-class generalized Jackson network, which is a key ingredient in justifying the interchange of limits for that model; refer to [6] (the inequalities in (25,28), in the proof of Theorem 3.3 in particular). Write the "free" process" in (25), $\hat{X}^k(t)$, as $\hat{X}^k(t) = \hat{A}^k(t) + \theta^k t$, where $\hat{A}^k(t)$ is approximately a driftless Brownian motion and $\theta^k := k(\rho^k - \rho) \to \theta$. Then, similar to the case of generalized Jackson network, the Lipschitz continuity, should it hold in the resource-sharing network, would require that the distance between the workload process in the pre-limit network $\hat{W}^k(t)$ and its associated fluid model $\hat{w}^k(t)$ be dominated by the distance between their corresponding free processes $\hat{X}^k(t)$ and $\theta^k t$. (Ignore the initial residuals for simplicity.) The latter distance being $|\hat{A}^k(t)|$, this would lead to (provided $\hat{W}^k(0) = \hat{w}^k(0)$):

$$|\hat{W}^k(t) - \hat{w}^k(t)| \leq \kappa |\hat{A}^k(t)|,$$

where κ a constant. As $\hat{w}^k(t) = 0$ when t is sufficiently large (and independent of k, guaranteed by uniform stability), we have

$$|\hat{W}^k(t)| \leq \kappa |\hat{A}^k(t)|,$$

where the p-th moment of the bound on the right hand side grows in the order of $t^{p/2}$. However, in most multi-class networks, Lipschitz continuity will not hold. Thus, in contrast, the bounded workload condition proposed above stipulates,

$$|\hat{W}^k(t)| \le \kappa |\hat{W}^k(0)| + \kappa \sup_{0\le s\le t} |\hat{A}^k(s) + \theta^k s|,$$

and we will show shortly in the next lemma that the p-th moment of this bound grows in the order of t^p. Viewed this way, the bounded workload condition can be regarded as a relaxed version of Lipschitz continuity; and in the rest of this section, we show the interchange of limits can be justified under this relaxed condition.

4.1. Key Estimates We start with the following lemma, which summarizes some of the direct implications of the bounded workload condition. The proof of the lemma is deferred to Appendix D.1.

Lemma 9 Assume the bounded workload condition in (72).
(a) (Bounded p-th Moment) Consider the sequence $\hat{\Xi}^k(t)$, with $|\hat{\Xi}^k(0)| \le M$ for some $M > 0$. The following holds for some constant κ,

$$\mathsf{E} \sup_{0\le s\le t} |\hat{W}^k(s)|^p \le \kappa (M^p + 1 + t^p); \tag{73}$$

and consequently (redefining κ), for any $0 \le q \le p$,

$$\mathsf{E} \sup_{0\le s\le t} |\hat{W}^k(s)|^q \le \kappa (M^q + 1 + t^q). \tag{74}$$

(b) (Uniform Integrability) Let $\{m_i; i = 1, 2, \cdots\}$ be a sequence of number such that $m_i \to \infty$ as $i \to \infty$; and let $\{x^i \in \mathcal{X}; i = 1, 2, \cdots\}$ be a sequence of initial states such that $|x^i| \le m_i$ for all i. Then, for any given $t \ge 0$, and a fixed, sufficiently large k, $\{|\hat{W}^k(m_i t; x^i)/m_i|^p\}$ is uniformly integrable (w.r.t. i).
(c) (Uniform Integrability) Let $\{m_k\}$ be a sequence of numbers such that $m_k \to \infty$ as $k \to \infty$, and assume that the sequence of initial states satisfies $|\hat{\Xi}^k(0)| \le m_k$. Then, $\{|\hat{W}^k(m_k t)/m_k|^p\}$ is uniformly integrable (w.r.t. k).

With the uniform stability and integrability results (Theorem 7(a) and Lemma 9), we are ready to establish the uniform p-th moment stability (Proposition 11) and holds the key to establishing the tightness of the stationary distributions associated with $\{\hat{W}^k(t)\}$. But first, we present a lemma, which provides key insights to this moment stability property, and serve as intermediate steps in the proof of the property as well.

Lemma 10 Under the usual traffic condition in (67), there exists a time t_0 such that the following conclusions hold.
(a) Let $\{m_i; i = 1, 2, \cdots\}$ be a sequence of number such that $m_i \to \infty$ as $i \to \infty$; and let $\{x^i \in \mathcal{X}; i = 1, 2, \cdots\}$ be a sequence of initial states such that $|x^i| \le m_i$ for all i. Then, for any sufficiently large k, the following holds (with probablity one), as $i \to \infty$,

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

(b) Let $\{m_k\}$ be a sequence of numbers such $m_k \to \infty$ as $k \to \infty$; and assume that the sequence of initial states $\{\hat{\Xi}^k(0)\}$ satisfies $|\hat{\Xi}^k(0)| \le m_k$. Then, the following holds (with probability one) as $k \to \infty$,

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

Suppose furthermore the bounded workload condition is satisfied. Then, the followings also hold.
(c) Assume $\{m_i\}$ and $\{x^i\}$ as in conclusion (a). Then, the following holds for sufficiently large k,

$$\lim_{i\to\infty} \mathsf{E}\frac{1}{m_i^p}\left|\hat{W}^k(m_i t; x^i)\right|^p = 0 \qquad \text{for } t \ge t_0.$$

(d) Assume $\{m_k\}$ and $\{\hat{\Xi}^k(0)\}$ as in conclusion (b). Then, the following holds,

$$\lim_{k\to\infty} \mathsf{E}\frac{1}{m_k^p}\left|\hat{W}^k(m_k t)\right|^p = 0 \qquad \text{for } t \ge t_0.$$

Parts (a) and (b) of the above lemma can be established by applying the uniform stability in Theorem 7(a) to the limits in Lemma 5 and Proposition 3(b), respectively. For part (c), note that the interchange of the expectation and the limit is justified by the uniform integrability in Lemma 9 (b), and then the conclusion follows from part (a). Part (d) is similarly proved by invoking Lemma 9 (c), along with part (b). The full proof of parts (a,c,d) is omitted, and that of part (b) is provided in Appendix D.2 for reference.

Proposition 11 (Uniform p-th moment stability) Under the usual traffic condition in (67) and the bounded workload condition in (72), there exists a time t_0 and a sufficiently large index k_0 such that the following holds for all $t \ge t_0$,

$$\lim_{|x|\to\infty} \sup_{k\ge k_0} \mathsf{E}\frac{1}{|x|^p}\left|\hat{W}^k(|x|t; x)\right|^p = 0. \tag{75}$$

Proof. Let t_0 be the same as in Lemma 10, and consider any time $t \ge t_0$. Suppose (75) is not true; then, there exists an $\epsilon_0 > 0$ and a sequence of initial states $\{x^i \in \mathcal{X} : i = 1, 2, \cdots\}$ satisfying $\lim_{i\to\infty} |x^i| = \infty$ such that

$$\sup_k \mathsf{E}\frac{1}{|x^i|^p}\left|\hat{W}^k(|x^i|t; x^i)\right|^p > 2\epsilon_0. \tag{76}$$

Corresponding to each x^i, choose an index in the sequence k, denoted by k_i, such that

$$\mathsf{E}\frac{1}{|x^i|^p}\left|\hat{W}^{k_i}(|x^i|t; x^i)\right|^p > \epsilon_0. \tag{77}$$

We claim that $\{k_i\}$ cannot be bounded. Otherwise, at least an index, say k', repeats in the sequence for infinite times; and clearly, this contradicts to Lemma 10(c). Without lost of generality, assume $k_i \to \infty$ as $i \to \infty$. Then, the bound in (77) contradicts to Lemma 10(d). □

It would be useful to note here that our approach is built upon various pathwise stability results leading to the *pathwise convergence* in Lemma 10(a,b). However, the uniform p-th moment stability (a *moment convergence*, as stated in Proposition 11 above), is required to prove the tightness and to justify the interchange of limits, the main results in the next subsection. To bridge the gap between these two modes of convergence, the uniform integrability stated in Lemma 9(b,c) is required; and that's why we introduced earlier the bounded workload condition, a pathwise condition that generates a growth in the power of p as described in Lemma 9(a) and thereafter ensures the needed uniform integrability.

4.2. Tightness and Interchange of Limits First, the estimate in the following proposition connects the key properties established in the previous subsection, specifically the bounded p-th moment and the uniform p-th moment stability of workload processes, to the tightness and the convergence of the stationary distributions in question. The estimate is concerned with a return time: For any time length $\delta > 0$ and any compact set $C \subset \mathcal{X}$, denote the return time to C after an initial period of δ as,

$$\tau_C^k(\delta) = \inf\{t \ge \delta : \hat{\Xi}^k(t) \in C\}.$$

Proposition 12 Under the usual traffic condition in (67) and the bounded workload condition in (72), there exist positive constants κ and δ, and a compact set $C \subset \mathcal{X}$ such that the following bound holds for any initial state $x \in \mathcal{X}$,

$$\sup_k \mathsf{E} \int_0^{\tau_C^k(\delta)} (1 + |\hat{\Xi}^k(t;x)|^{p-1}) dt \le \kappa(1 + |x|^p). \tag{78}$$

Given Lemma 9(a) and Proposition 11, the proof of the above proposition is a slight modification of those for Theorem 3.4 of [6] and Proposition 5.3 of [15]; hence, it is omitted.

Next, recall that in Theorem 7, $\hat{\pi}^k$ denotes the unique stationary distribution of $\hat{\Xi}^k(t)$. Let $\hat{\pi}_1^k$ denote the stationary distribution of $\hat{W}^k(t)$, the first component of $\hat{\Xi}^k(t)$; and $\hat{\pi}_1$ be that of $\hat{W}(t)$.

Proposition 13 Under the usual traffic condition in (67) and the bounded workload condition in (72), the sequence of stationary distributions, $\{\hat{\pi}^k\}$, is tight on $\mathcal{X}$. Furthermore, we have $\sup_k \mathsf{E}_{\hat{\pi}^k} |\hat{\Xi}^k(0)|^{p-1} < \infty$ (which strengthens the conclusion in (62)).

Proof. First, the following result can be found from the proof for Theorem 3.2 of [6] along with Theorem 3.5 of [6] (also Proposition 5.4 of [15]): For any given constant $\delta > 0$, and compact set $C \subset \mathcal{X}$, define

$$V_k(x) = \mathsf{E} \int_0^{\tau_C^k(\delta)} f(\hat{\Xi}^k(t;x)) dt, \qquad x \in \mathcal{X},$$

where $f(x) \ge 0$, $x \in \mathcal{X}$, is any given function such that the above expectation exists. Suppose $\sup_k V_k(x)$ is finite for all x and uniformly bounded on C. Then, there exists a constant κ such that the following bound holds for all k:

$$\mathsf{E}_{\hat{\pi}^k} f(\hat{\Xi}^k(0)) = \int_{\mathcal{X}} f(x) \hat{\pi}^k(dx) \le \kappa. \tag{79}$$

Now, let $f(x) = 1 + |x|^{p-1}$. Then, the finiteness and uniform bound conditions on $V_k(x)$ in the above can be justified by Proposition 12, with δ and C also specified in Proposition 12. From

(79), we have the following uniform bound for the stationary moments of the networks: for some constant κ,

$$\mathsf{E}_{\hat{\pi}^k}|\hat{\Xi}^k(0)|^{p-1} \leq \kappa \quad \text{for all } k.$$

The above bound (with $p-1 \geq 1$) also implies the tightness of the sequence of stationary distributions $\{\hat{\pi}^k\}$. □

Finally, we establish the interchange of limits in the following theorem.

Theorem 14 Under the usual traffic condition in (67) and the bounded workload condition in (72), the following weak convergence of stationary distributions holds,

$$\hat{\pi}_1^k \Rightarrow \hat{\pi}_1, \quad \text{as } k \to \infty. \tag{80}$$

In particular, since $\hat{\pi}_1$ is 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 (sides III and IV) is valid. Furthermore, for any $m \in [0, p-1)$,

$$\mathsf{E}_{\hat{\pi}_1^k}|\hat{W}^k(0)|^m \to \mathsf{E}_{\hat{\pi}_1}|\hat{W}(0)|^m, \quad \text{as } k \to \infty. \tag{81}$$

Proof. Once Proposition 13 is proven, the weak convergence, $\hat{\pi}_1^k \Rightarrow \hat{\pi}_1$, follows from a rather standard argument, similar to the one in Gamarnik and Zeevi [23] (see also [6, 26, 34, 35]), which we outline below for completeness.

First, recall that Theorem 7 (b) (or Theorem 8 (c)) guarantees the existence of the stationary distribution $\hat{\pi}^k$ for sufficiently large k. Following Proposition 13, the sequence $\{\hat{\pi}^k\}$ is tight, which implies that $\{\hat{\pi}_1^k\}$ is also tight ([1], page 65). This, in turn, implies that for any subsequence of k, there exists a further subsequence $\mathcal{K}_1$ such that $\{\hat{\pi}_1^k\}$ converges weakly along $\mathcal{K}_1$ to a limiting distribution, which we (tentatively) denote as $\tilde{\pi}_1$.

Next, we initialize the process $\hat{\Xi}^k(t)$ in its stationary distribution $\hat{\pi}^k$; and since $\{\hat{\pi}^k\}$ is tight, Proposition 2 can be applied to the sequence of networks in $\mathcal{K}_1$. Hence, there exists a subsequence $\mathcal{K}_2 \subset \mathcal{K}_1$ such that $\{\hat{W}^k(t_0^k + t); k \in \mathcal{K}_2\}$ (with t_0^k specified in Proposition 2, and $t \geq 0$ being any given time) converges weakly to a limit $\hat{W}(t)$, as characterized in (28-33). Due to our choice of initialization, $\hat{W}^k(t_0^k + t)$ is equal in distribution to $\hat{W}^k(0)$, for each $k \in \mathcal{K}_2$. 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 $\hat{\pi}_1$ guaranteed by Theorem 4 (or Theorem 8(b)). In summary, we can conclude that $\hat{\pi}_1$ is the weak limit of any convergent subsequence of $\{\hat{\pi}_1^k\}$. Therefore, the full sequence $\{\hat{\pi}_1^k\}$ must converge weakly to $\hat{\pi}_1$. (Note that we cannot assume, a priori, that the limiting distribution $\tilde{\pi}_1$ takes on values solely in the fixed-point state space $\mathcal{W}$. Therefore, Theorem 1 does not apply directly to the sequence of networks initialized in $\{\hat{\pi}^k\}$. Proposition 2 fills the gap instead.)

Lastly, the convergence in (81) is a direct consequence of the convergence in (80) and the "furthermore" part in Proposition 13. □

5. Verifying the Bounded Workload Condition

With the results in the last section, to justify the interchange of limits is reduced to verifying the bounded workload condition. Here we first give three illustrative examples, and then present a sufficient condition for bounded workload. Among the three examples, the first one has a reflection matrix that is an M-matrix and the DCP satisfies the complementarity condition; the second example does not have an M-matrix, and the third one fails the complementarity. The bounded workload condition, however, can be verified in all three cases.

5.1. M-Matrix and Complementarity Consider the sequence of resource-sharing networks in Theorem 1, with two additional conditions.

- *M-matrix*: G^TBG is an M-matrix (e.g., [2]); in particular, it is a matrix with positive diagonal entries and non-positive off-diagonal entries and its inverse is a non-negative matrix.
- *Complementarity*: In addition to the specifications in (23-27), the complementarity condition as described in (31) also holds for the k-th (pre-limit) network. Specifically, the k-th network satisfies the following DCP (given the transformation $\hat{W}_G^k(t) = G^T\hat{W}^k(t)$),

$$\hat{W}_G^k(t) = \hat{W}_G^k(0) + G^T\hat{X}^k(t) + G^TBG\hat{Y}^k(t) \quad \text{for } t \geq 0; \tag{82}$$

$$\hat{Y}_\ell^k(t) \text{ is non-decreasing in } t \geq 0,\ \hat{Y}_\ell^k(0) = 0, \quad \ell \in \mathcal{L}; \tag{83}$$

$$\hat{Y}_\ell^k(t) \text{ cannot increase at } t, \text{ if } \hat{W}_{G,\ell}^k(t) > 0, \quad \ell \in \mathcal{L}. \tag{84}$$

If we add to the above DCP the non-negativity condition:

$$\hat{W}_G^k(t) \geq 0, \quad \text{for } t \geq 0; \tag{85}$$

then, it becomes the standard Skorohod problem that defines the reflected Brownian motion (RBM) with the reflection matrix G^TBG being an M-matrix, and the free process (un-reflected Brownian motion) being $(G^T\hat{W}^k(0) + G^T\hat{X}^k(t))$. It is known (e.g., Theorem 7.2 of [8]) that this Skorohod problem has a unique solution, which we denote as $\hat{Y}^{*,k}$ and $\hat{W}^{*,k}$.

However, the non-negativity in (85) need not hold for the DCP in (82-84). In other words, when the free process drags $\hat{W}_G^k$ down to the negative region, the reflection process $\hat{Y}^k$ increases, but not enough to keep $\hat{W}_G^k$ above zero. This is in contrast with $\hat{Y}^{*,k}$, which increases hard enough to maintain the non-negativity of $\hat{W}_G^{*,k}$. Consequently, we have

$$\hat{Y}^k(t) \leq \hat{Y}^{*,k}(t), \quad \text{for } t \geq 0. \tag{86}$$

(Refer to the proof of Lemma 2.1 in [7].) Since $\hat{W}_G^{k,*}$ and $\hat{W}_G^k$ are driven by the same free process, we have

$$\hat{W}_G^{*,k}(t) - \hat{W}_G^k(t) = G^TBG[\hat{Y}^{*,k}(t) - \hat{Y}^k(t)] \tag{87}$$

Multiplying both sides by $ABA^T = (G^TBG)^{-1}$, and taking into account $(ABA^T)G^T = A$, we have

$$A\hat{W}^{*,k}(t) - A\hat{W}^k(t) = \hat{Y}^{*,k}(t) - \hat{Y}^k(t) \geq 0. \tag{88}$$

Applying the oscillation inequality in Lemma 21, we can dominate the oscillation of $\hat{Y}^{*,k}(t)$ by that of the free process $(G^T\hat{W}^k(0) + G^T\hat{X}^k(t))$. (Alternatively, we can apply the stronger Lipschitz continuity of the Skorohod mapping for RBM; refer to, e.g., Theorem 7.2 of [8].) Hence, we have

$$\sup_{0\leq s\leq t} |A\hat{W}^k(s)| \leq \sup_{0\leq s\leq t} |A\hat{W}^{*,k}(s)| \leq \kappa' \left(|\hat{W}^k(0)| + \sup_{0\leq s\leq t} |\hat{X}^k(s)| \right), \quad t \geq 0; \tag{89}$$

which implies the bounded workload condition in (72) for the constant $\kappa = \kappa'/a_{\min}$, where $a_{\min} = \min\{a_{\ell r} : a_{\ell r} > 0\}$.

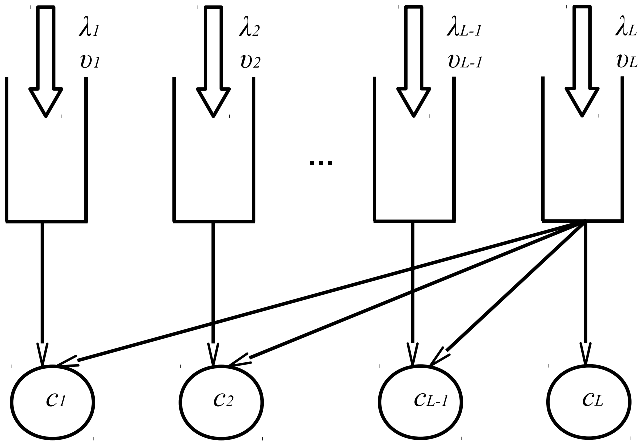

Figure 4. A network in which one class occupies all servers

As a concrete example, the above two conditions hold for the network in Figure 4. It consists of a set of L servers and the same number of job classes, $R = L$; each of the first $L-1$ class uses a single server, while the last class L requires simultaneous occupancy of all servers. Note, the capacity of every server ℓ, with the exception of the last one, is shared among two classes $r = \ell$ and L. The last server serves the last class L only (which also uses every other server). The heavy traffic condition reads: $\rho_L = c_L$ and $\rho_\ell + \rho_L = c_\ell$ for $\ell = 1, \cdots, L-1$; i.e., all servers are bottlenecks.

The incidence matrix of this network example is

$$A = \begin{pmatrix} 1 & 0 & & 0 & 1 \\ 0 & 1 & & 0 & 1 \\ & & \ddots & & \\ 0 & 0 & & 1 & 1 \\ 0 & 0 & & 0 & 1 \end{pmatrix}_{L \times L}.$$

It is straightforward to derive the following,

$$ABA^T = \begin{pmatrix} b_1 + b_L & b_L & & b_L & b_L \\ b_L & b_2 + b_L & & b_L & b_L \\ & & \ddots & & \\ b_L & b_L & & b_{L-1} + b_L & b_L \\ b_L & b_L & & b_{L-1} & +b_L \end{pmatrix}, \quad G = \begin{pmatrix} \frac{1}{b_1} & 0 & & 0 & -\frac{1}{b_1} \\ 0 & \frac{1}{b_2} & & 0 & -\frac{1}{b_2} \\ & & \ddots & & \\ 0 & 0 & & \frac{1}{b_{L-1}} & -\frac{1}{b_{L-1}} \\ 0 & 0 & & 0 & \frac{1}{b_L} \end{pmatrix},$$

$$G^T B G = (ABA^T)^{-1} = \begin{pmatrix} \frac{1}{b_1} & 0 & & 0 & -\frac{1}{b_1} \\ 0 & \frac{1}{b_2} & & 0 & -\frac{1}{b_2} \\ & & \ddots & & \\ 0 & 0 & & \frac{1}{b_{L-1}} & -\frac{1}{b_{L-1}} \\ -\frac{1}{b_1} & -\frac{1}{b_2} & & -\frac{1}{b_{L-1}} & \sum_{\ell=1}^{L} \frac{1}{b_\ell} \end{pmatrix}.$$

Clearly, G^TBG is an M-matrix, since it has non-positive off-diagonal elements and is inverse-positive, i.e., $(G^TBG)^{-1} = ABA^T \geq 0$.

Given the matrix G, the complementarity condition in (84) reads:

$$\text{if } \hat{W}^k_\ell(t) > 0, \text{ then server } \ell \text{ is fully utilized}, \quad \ell = 1, \cdots, L-1; \quad \text{and} \tag{90}$$

$$\text{if } \frac{\hat{W}^k_L(t)}{b_L} > \sum_{\ell=1}^{L-1} \frac{\hat{W}^k_\ell(t)}{b_\ell}, \text{ then server } L \text{ is fully utilized.} \tag{91}$$

The first condition (90) is obvious. So, consider server L and the case that $\hat{W}^k_L(t)/b_L > \sum_{r=1}^{L-1} \hat{W}^k_r(t)/b_r$. Without loss of generality, consider the case of $\hat{W}^k_r(t) > 0$, $r = 1, \cdots, L-1$. Suppose to the contrary of (91), server L is not fully utilized, or $\Lambda_L(\hat{N}^k(t)) < c_L = \rho_L$, which also implies $\Lambda_r(\hat{N}^k(t)) > \rho_r$ for $r = 1, \cdots, L-1$. From the KKT condition of the optimization problem in (3), with n replaced by $\hat{N}^k(t)$ (recall $\hat{W}^k_r(t) = \nu_r \hat{N}^k_r(t)$), we find that

$$\frac{\beta_r \hat{N}^k_r(t)}{\Lambda_r(\hat{N}^k(t))} = \eta_\ell, \quad r = \ell = 1, \cdots, L-1, \quad \text{and}$$
$$\frac{\beta_L \hat{N}^k_L(t)}{\Lambda_L(\hat{N}^k(t))} = \sum_{\ell=1}^{L} \eta_\ell,$$

where η_ℓ denotes the shadow price of the corresponding server. Note that $\eta_L = 0$, since server L is under-utilized (as assumed). Consequently,

$$\frac{\rho_L}{\Lambda_L(\hat{N}^k(t))} \frac{\hat{W}^k_L(t)}{b_L} = \sum_{r=1}^{L-1} \frac{\rho_r}{\Lambda_r(\hat{N}^k(t))} \frac{\hat{W}^k_r(t)}{b_r} \leq \sum_{r=1}^{L-1} \frac{\hat{W}^k_r(t)}{b_r} < \frac{\hat{W}^k_L(t)}{b_L}.$$

The above implies $\Lambda_L(\hat{N}^k(t)) > \rho_L = c_L$, which violates the capacity constraint of server L.

In summary, the network in Figure 4 meets both requirements, M-matrix and complementarity, and therefore satisfies the bounded workload condition as we have established earlier.

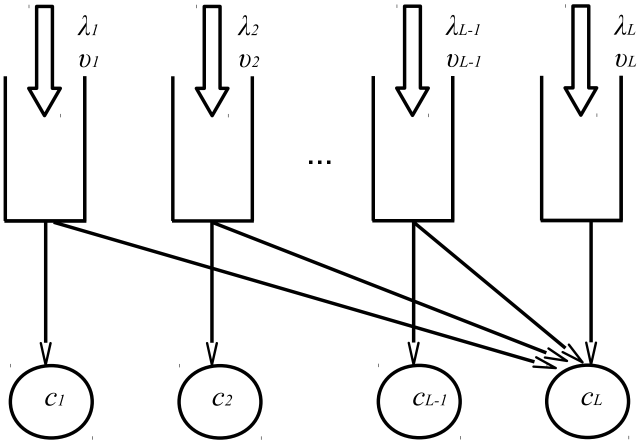

Figure 5. A network in which one server is shared by all classes

5.2. Without M-Matrix Consider the network in Figure 5, with server L shared by all job classes, whereas every other server serves one class only. In this case, the incidence matrix A is the transpose of its counterpart in the last example:

$$A = \begin{pmatrix} 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & 0 \\ 1 & 1 & & 1 & 1 \end{pmatrix}_{L \times L}.$$

We have,

$$ABA^T = \begin{pmatrix} b_1 & 0 & & 0 & b_1 \\ 0 & b_2 & & 0 & b_2 \\ & & \ddots & & \\ 0 & 0 & & b_{L-1} & b_{L-1} \\ b_1 & b_2 & & b_{L-1} & \sum_{\ell=1}^{L} b_\ell \end{pmatrix}, \quad G = \begin{pmatrix} \frac{1}{b_1} & 0 & & 0 & 0 \\ 0 & \frac{1}{b_2} & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & \frac{1}{b_{L-1}} & 0 \\ -\frac{1}{b_L} & -\frac{1}{b_L} & & -\frac{1}{b_L} & \frac{1}{b_L} \end{pmatrix},$$

$$G^T BG = \begin{pmatrix} \frac{1}{b_1} + \frac{1}{b_L} & \frac{1}{b_L} & & \frac{1}{b_L} & -\frac{1}{b_L} \\ \frac{1}{b_L} & \frac{1}{b_2} + \frac{1}{b_L} & & \frac{1}{b_L} & -\frac{1}{b_L} \\ & & \ddots & & \\ \frac{1}{b_L} & \frac{1}{b_L} & & \frac{1}{b_{L-1}} + \frac{1}{b_L} & -\frac{1}{b_L} \\ -\frac{1}{b_L} & -\frac{1}{b_L} & & -\frac{1}{b_L} & \frac{1}{b_L} \end{pmatrix}.$$

Thus, $G^T BG$ is *not* an M-matrix (when $L \geq 3$), since it has positive off-diagonal elements. Nevertheless, the bounded workload condition still holds, as will be shown below.

Consider any k, and fix any time $t > 0$. We first estimate the class-L workload. Suppose $\hat{W}_L^k(t) > 0$ without loss of generality. Let $t_0 \geq 0$ be the smallest time such that

$$\hat{W}_L^k(s) > 0, \quad \text{for all } s \in (t_0, t].$$

Observe that at any time $s \in (t_0, t]$, the server L is fully occupied, i.e.,

$$\sum_{r=1}^{L} \Lambda_r(\hat{N}^k(s)) = c_L.$$

Then, we have

$$\Lambda_L(\hat{N}^k(s)) = c_L - \sum_{r=1}^{L-1} \Lambda_r(\hat{N}^k(s)) \geq c_L - \sum_{\ell=1}^{L-1} c_\ell = \rho_L,$$

where the inequality is because the allocations to class r, $r = 1, \cdots, L-1$, are limited to the capacities of server ℓ ($\ell = r$). Hence, regarding the last items in (23), we have from (24),

$$k(\rho_L t - \tilde{D}_L^k(t)) - k(\rho_L t_0 - \tilde{D}_L^k(t_0)) = k \int_{t_0}^{t} (\rho_L - \Lambda_L(\hat{N}^k(s)))ds \leq 0.$$

Consequently, we have from (23),

$$\hat{W}_L^k(t)-\hat{W}_L^k(t_0)=\hat{X}_L^k(t)-\hat{X}_L^k(t_0)+k(\rho_L t-\tilde{D}_L^k(t))-k(\rho_L t_0-\tilde{D}_L^k(t_0))\le \hat{X}_L^k(t)-\hat{X}_L^k(t_0).$$

And, no matter whether $t_0=0$ or $t_0>0$, the above implies

$$\hat{W}_L^k(t)\le \hat{W}_L^k(0)+2\sup_{0\le s\le t}|\hat{X}_L^k(s)|. \tag{92}$$

Next, we estimate the class-ℓ workload for $\ell=1,\cdots,L-1$. Suppose without loss of generality

$$\hat{W}_\ell^k(t)/b_\ell>\hat{W}_L^k(t)/b_L; \tag{93}$$

otherwise, $\hat{W}_\ell^k(t)$ can be bounded through (92). Let $t_0\ge 0$ be the smallest time such that

$$\frac{\hat{W}_\ell^k(s)}{b_\ell}>\frac{\hat{W}_L^k(s)}{b_L},\quad \text{for all } s\in(t_0,t]. \tag{94}$$

Under the above condition, server ℓ must be fully occupied:

$$\Lambda_\ell(\hat{N}^k(s))=\rho_\ell,\quad \text{for all } s\in(t_0,t]. \tag{95}$$

Suppose to the contrary, $\Lambda_\ell(\hat{N}^k(s'))<\rho_\ell$ for a time $s'\in(t_0,t]$. From the KKT condition of the optimization problem in (3), with n replaced by $\hat{N}^k(s')$, we find that

$$\frac{\beta_\ell \hat{N}_\ell^k(s')}{\Lambda_\ell(\hat{N}^k(s))}=\eta_\ell+\eta_L=\eta_L=\frac{\beta_L\hat{N}_L^k(s')}{\Lambda_L(\hat{N}^k(s))},$$

where η_ℓ and η_L are the shadow prices of servers ℓ and L, respectively; and we have $\eta_\ell=0$ as server ℓ is assumed to be under-utilized. Rewrite the above as follows:

$$\frac{\rho_\ell}{\Lambda_\ell(\hat{N}^k(s))}\frac{\hat{W}_\ell^k(s')}{b_\ell}=\frac{\rho_L}{\Lambda_L(\hat{N}^k(s))}\frac{\hat{W}_L^k(s')}{b_L}.$$

This equality, along with the condition in (94) and the contradictory assumption, implies $\Lambda_L(\hat{N}^k(s'))<\rho_L$. That is, both servers ℓ and L are under-utilized at time s'. Such an allocation certainly cannot maximize the utility objective function in (3).

Using the property in (95), we have

$$\hat{W}_\ell^k(t)-\hat{W}_\ell^k(t_0)=\hat{X}_\ell^k(t)-\hat{X}_\ell^k(t_0)+k\int_{t_0}^t(\rho_\ell-\Lambda_\ell(\hat{N}^k(s)))ds=\hat{X}_\ell^k(t)-\hat{X}_\ell^k(t_0).$$

Consequently, no matter whether $t_0=0$ or $t_0>0$, we have

$$\begin{aligned}\hat{W}_\ell^k(t)&\le (b_\ell/b_L)\hat{W}_L^k(t_0)+\hat{W}_\ell^k(0)+\hat{X}_\ell^k(t)-\hat{X}_\ell^k(t_0)\\&\le (b_\ell/b_L)(\hat{W}_L^k(0)+2\sup_{0\le s\le t}|\hat{X}_L^k(s)|)+\hat{W}_\ell^k(0)+2\sup_{0\le s\le t}|\hat{X}_\ell^k(s)|.\end{aligned} \tag{96}$$

Finally, as t is arbitrary, the estimates in (92, 96) imply that the bounded workload condition in (72) must be satisfied.

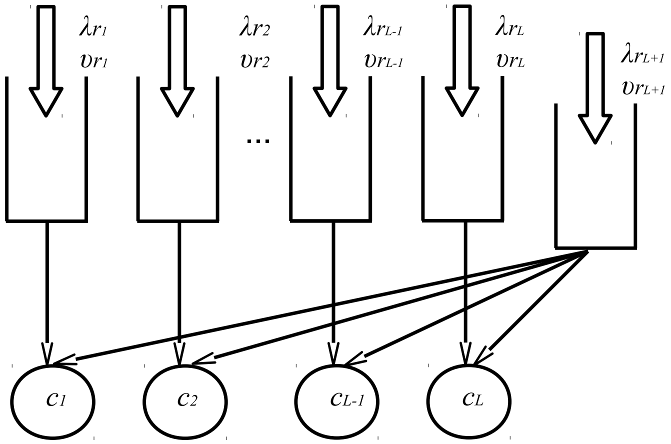

FIGURE 6. A symmetric variation of the network in Figure 4

5.3. Without Complementarity Consider the network in Figure 6. It is a model studied by others in the literature (and sometimes referred to as a "linear network") with wide-ranging applications, from the Internet protocol to road traffic control (e.g., [43, 32, 39]). It is, in fact, a symmetric variation of our first example, the network in Figure 4: with the addition of class $L+1$, which is the class that occupies all L servers, now every server ℓ serves exactly two classes, class ℓ and class L. (In the first example server L serves one class only.) The incidence matrix is no longer a square matrix:

$$A = \begin{pmatrix} 1\,0 & & 0\,0\,1 \\ 0\,1 & & 0\,0\,1 \\ & \ddots & \\ 0\,0 & & 1\,0\,1 \\ 0\,0 & & 0\,1\,1 \end{pmatrix}_{L\times(L+1)} .$$

To avoid tedious algebra, consider the following parameters: $\lambda_r = 1$, $\nu_r^k = \nu_r = 1$, $\rho_r = 1$, $\beta_r = 1$, $c_\ell = 2$, $B = I$. Then, the matrices G, H and $G^T BG$ are as follows:

$$G = \frac{1}{L+1}\begin{pmatrix} L & -1 & & -1 \\ -1 & L & & -1 \\ & & \ddots & \\ -1 & -1 & \cdots & L \\ 1 & 1 & 1 & 1 \end{pmatrix}_{(L+1)\times L}, \quad H = \frac{1}{\sqrt{L+1}}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ -1 \end{pmatrix}_{(L+1)\times 1},$$

$$G^T BG = \frac{1}{L+1}\begin{pmatrix} L & -1 & & -1 \\ -1 & L & & -1 \\ & & \ddots & \\ -1 & -1 & & L \end{pmatrix}_{L\times L} .$$

The optimization problem in (3) for the proportional fair allocation now reads:

$$\max_{\gamma} \sum_{r=1}^{L+1} n_r \log(\gamma_r), \tag{97}$$

$$\text{s.t.} \quad \gamma_r + \gamma_{L+1} \le c_\ell, \quad r = \ell = 1, \cdots, L. \tag{98}$$

Solving the above problem yields the following allocation,

$$\Lambda_{L+1}(n)=\frac{w_{L+1}}{\sum_{\ell=1}^{L+1}w_\ell}c_\ell,\quad \text{for } w\neq 0, \tag{99}$$

$$\Lambda_\ell(n)=c_\ell-\Lambda_{L+1}(n),\quad \text{for } w_\ell>0,\quad \text{for } \ell=1,\cdots,L. \tag{100}$$

Consider a state w satisfying $w_1=0$ and $w_{L+1}>\sum_{\ell=2}^{L}w_\ell>0$. Hence, we have $g_1^T w=Lw_1+w_{L+1}-\sum_{\ell=2}^{L}w_\ell>0$. However, from (99) and (100), we have

$$\Lambda_\ell(n)+\Lambda_{L+1}(n)=0+\frac{w_{L+1}}{\sum_{\ell=1}^{L+1}w_\ell}c_\ell<c_\ell;$$

that is, link 1 is not fully occupied, even though $g_1 w>0$. Therefore, this network does not satisfy the complementarity condition of the first example. Nevertheless, we can still verify the bounded workload condition as follows.

First, from (17, 18) and (23), we have

$$\begin{aligned} H^T\hat{W}^k(t) &= H^T\hat{W}^k(0)+H^T\hat{X}^k(t)+\hat{Z}^k(t)\\ &= H^T\hat{W}^k(0)+H^T\hat{X}^k(t)+\int_0^t H^T[\rho-\Lambda(\hat{N}^k(s))]ds. \end{aligned} \tag{101}$$

Using (99) and (100), we can show that

$$[H^T\hat{W}^k(t)]\cdot[H^T(\rho-\Lambda(\hat{N}^k(s)))]=-\sqrt{L+1}\frac{\left(\sum_{\ell=1}^{L}\hat{W}_\ell^k(t)-\hat{W}_{L+1}^k(t)\right)^2}{\sum_{\ell=1}^{L+1}\hat{W}_\ell^k(t)}\leq 0.$$

In view of (101), the above implies that $\hat{Z}^k(t)$ cannot increase (resp. decrease) at time t if $H^T\hat{W}^k(t)>0$ (resp. <0). Hence, intuitively, the deviation of $H^T\hat{W}^k(t)$ from zero is mitigated by $\hat{Z}^k(t)$, and consequently cannot exceed the oscillation of $H^T\hat{X}^k(t)$.

To formalize the above intuition, we consider any fixed time $t>0$ for the moment, and consider the case of $H^T\hat{W}^k(t)>0$ first. Let $t_0\geq 0$ be the smallest time such that

$$H^T\hat{W}^k(s)>0,\quad \text{for all } s\in(t_0,t].$$

This implies

$$\hat{Z}^k(t)-\hat{Z}^k(t_0)=\int_{t_0}^t H^T(\rho-\Lambda(\hat{N}^k(s)))ds\leq 0.$$

Then, we have

$$H^T\hat{W}^k(t)-H^T\hat{W}^k(t_0-)=H^T\hat{X}^k(t)-H^T\hat{X}^k(t_0-)+\hat{Z}^k(t)-\hat{Z}^k(t_0)\leq H^T\hat{X}^k(t)-H^T\hat{X}^k(t_0-).$$

Here, in the case of $t_0=0$, we understand $\hat{W}^k(t_0-)$ and $\hat{X}^k(t_0-)$ as $\hat{W}^k(0)$ and $\hat{X}^k(0)$, respectively; and also note that $\hat{Z}^k(s)$ is continuous in time s. No matter whether $t_0=0$ ($\hat{W}^k(t_0-)=\hat{W}^k(0)$) or $t_0>0$ ($\hat{W}^k(t_0-)=0$), the above inequality implies

$$|H^T\hat{W}^k(t)|\leq|H^T\hat{W}^k(0)|+\mathsf{Osc}(H^T\hat{X}^k(\cdot),[0,t]).$$

For the other case that $H^T\hat{W}^k(t)<0$, we can derive the above estimate simliarly. As the time t is given arbitrarily, we have

$$\sup_{0\le s\le t}|H^T\hat{W}^k(s)|\le|H^T\hat{W}^k(0)|+\mathsf{Osc}(H^T\hat{X}^k(\cdot),[0,t]). \tag{102}$$

In addition, the above, along with (101), implies the following bound for $\hat{Z}^k(t)$, for some constant $\kappa_1>0$,

$$\begin{aligned}\sup_{0\le s\le t}|\hat{Z}^k(s)| &\le 2\left(|H^T\hat{W}^k(0)|+\mathsf{Osc}(H^T\hat{X}^k(\cdot),[0,t])\right)\\ &\le \kappa_1\left(|\hat{W}^k(0)|+\sup_{0\le s\le t}|\hat{X}^k(s)|\right).\end{aligned} \tag{103}$$

Next, let $\hat{W}^k_{(L)}(t)$, $\hat{X}^k_{(L)}(t)$, G_L and H_L be the first L rows of $\hat{W}^k(t)$, $\hat{X}^k(t)$, G and H, respectively. It can be observed that the following relations hold,

$$\begin{aligned}&\hat{W}^k_{(L)}(t)=\hat{W}^k_{(L)}(0)+\hat{X}^k_{(L)}(t)+G_L\hat{Y}^k(t)+H_L\hat{Z}^k(t)\ge 0,\\ &\hat{Y}^k(t)\text{ is non-decreasing, }\hat{Y}^k(0)=0,\\ &\hat{W}^k_{(L)}(t)^Td\hat{Y}^k(t)=0.\end{aligned}$$

The last equation holds because server ℓ will be fully utilized when there are jobs present from its dedicated class $r=\ell$; refer to (100). Note that G_L is an M-matrix (and hence, a complete-S matrix); applying the oscillation inequality in Theorem 5.1 of [53] and then using the bound in (103), we have for some positive constants κ_2 and κ_3,

$$\begin{aligned}\sup_{0\le s\le t}|\hat{W}^k_{(L)}(s)| &\le |\hat{W}^k_{(L)}(0)|+\kappa_2\cdot\mathsf{Osc}\{\hat{X}^k_{(L)}(\cdot)+H_L\hat{Z}^k(\cdot),[0,t]\}\\ &\le \kappa_3\left(|\hat{W}^k(0)|+\sup_{0\le s\le t}|\hat{X}^k(s)|\right).\end{aligned} \tag{104}$$

Moreover, observe that, $\hat{W}^k_{L+1}=|\hat{W}^k_{(L)}|-\sqrt{L+1}H^T\hat{W}^k$. Then, from the bounds in (102) and (104), we have for some positive constant κ_4,

$$\sup_{0\le s\le t}|\hat{W}^k_{L+1}(s)|\le\kappa_4\left(|\hat{W}^k(0)|+\sup_{0\le s\le t}|\hat{X}^k(s)|\right). \tag{105}$$

From the bounds in (104) and (105), we know that for some positive constant κ, the bounded workload condition in (72) holds.

5.4. A Sufficient Condition for Bounded Workload It turns out there is a sufficient condition for bounded workload, which provides an alternative to proving the latter condition directly, as well as sheds more light to it. The sufficient condition takes the form of d^{fp}, the distance from the fixed-point state space:

$$\sup_{0\le s\le t^*}d^{fp}(\hat{W}^k(s))\le d^{fp}(\hat{W}^k(0))+\kappa^*\sup_{0\le s\le t^*}|\hat{X}^k(s)|,\quad\text{for any }t^*\ge 0, \tag{106}$$

Roughly speaking, the condition requires that the distance (of the workload) from the fixed-point state space be bounded by the initial distance plus the "free" process. Intuitively, once close to the fixed-point state space, the (pre-limit) network is required to stay within bounded distance from the diffusion limit; consequently, its performance (such as stability and moments) can be approximated by that of the latter, and thus justifying the interchange of limits. (In Eryilmaz and Srikant [20], a similar observation is made for a parallel server system and a wireless network model under under heavy traffic.)

Indeed, in two of the above examples verifying the bounded workload condition takes the form of bounding $d^{fp}(\hat{W}^k(t))$ in (22). Specifically, for the example in §5.2, we can write

$$d^{fp}(w) = \sum_{\ell=1}^{L-1} \left(\frac{w_L}{b_L} - \frac{w_\ell}{b_\ell} \right)^+,$$

and then bound the workload for each link ℓ such that $w_L/b_L > w_\ell/b_\ell$. For the example in §5.3, the key step turns out to be bounding of the (other) term of $d^{fp}(w)$, $h_m^T|\hat{W}^k(t)|$, as specified in (102).

To show that the inequality in (106) implies the bounded workload condition, consider any fixed time $t^* \geq 0$, and for convenience, denote $c' := \sup_{0\leq s\leq t^*} |\hat{W}^k(s)|$, and denote the right-hand-side of (106) as c''. We need to make use of Lemma 2 of Ye and Yao [60], reproduced in the Appendix as Lemma 19, where the parameters are chosen as $\kappa = 1$, $\epsilon := \epsilon'$ to be sufficiently small, and $\sigma := \sigma' (\leq \epsilon')$.

Case 1, $c''/c' > \sigma'$. This gives the bounded workload condition in (72).

Case 2, $c''/c' \leq \sigma'$. This along with the condition in (106) implies

$$d_{fp}\left(\frac{\hat{W}^k(t)}{c'} \right) \leq \frac{c''}{c'} \leq \sigma', \quad 0 \leq t \leq t^*. \tag{107}$$

Note that $|\hat{W}^k(t)/c'| \leq \kappa = 1$. According to Lemma 19, for any time $t \in [0, t^*]$, a link ℓ will be fully occupied (or $\hat{Y}_\ell^k(t)$ can not increase) if

$$g_\ell \hat{W}^k(t) \geq c'\epsilon'.$$

Moreover, from the definition in (22) and the inequality in (107), we can require that for $t \in [0, t^*]$,

$$g_\ell \hat{W}^k(t) \geq -c'\epsilon'.$$

Now, we can apply the oscillation inequality in Lemma 21 to the diffusion scaled system, with the time restricted to $[0, t^*]$ and ϵ being replaced by $c'\epsilon'$; hence, we have

$$\mathsf{Osc}(\hat{W}^k(\cdot), [s,t]) \leq \kappa_c (\mathsf{Osc}(\hat{X}^k(\cdot), [s,t]) + \epsilon' \sup_{0\leq s\leq t^*} |\hat{W}^k(s)|), \quad 0 \leq s \leq t \leq t^*,$$

which gives, for any time $t \in [0, t^*]$,

$$|\hat{W}^k(t)| \leq |\hat{W}^k(0)| + \mathsf{Osc}(\hat{W}^k(\cdot), [0, t^*]) \leq |\hat{W}^k(0)| + \kappa_c \left(2 \sup_{0\leq s\leq t^*} |\hat{X}^k(s)| + \epsilon' \sup_{0\leq s\leq t^*} |\hat{W}^k(s)| \right).$$

Note that ϵ' can be chosen such that $\kappa_c \epsilon' < 1$ from the beginning, and that the time t^* is arbitrarily given. Then, the above inequality implies the bounded workload condition.

6. Extension

Our approach outlined in §3∼§5 above can be readily extended to stochastic processing networks other than the resource-sharing networks that we have so far focused on. This should include the multi-class queueing networks studied in, for instance, [4, 9, 10, 54], and other stochastic processing networks as in [14, 50].

As an illustrative example, consider the two-station network depicted in Figure 7, first studied by Kumar and Seidman [41], and by Rybko and Stolyar [47] independently, referred to below as the KS-RS network. The diffusion limit (side I) for the KS-RS network has been established in Chen and Zhang [10], and extended to more general multi-class queueing network under priority disciplines in Chen and Zhang [10] and Chen and Ye [9]. Below, we spell out how our approach can be extended to establishing the other three sides in Figure 1 for the KS-RS network, in particular side IV—via verification of the bounded workload condition. (Note that results concerning sides II and III are available in the existing literature, and that side IV can be established by the approach of Gurvich [26] assuming sufficiently high moment conditions on the interarrival and service times.)

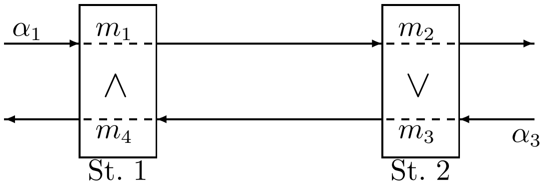

Figure 7. KS-RS network

6.1. KS-RS Network and its Diffusion Limit

The KS-RS network consists of two service stations, each with a single server and an infinite-size buffer. There are two external arrival streams, indexed as class-1 and class-3 jobs. A class-1 (class-3) job will turn into a class-2 (class-4) job after its service completion at station 1 (station 2), and then leave the system after its service completion at station 2 (station 1). Within the same class, jobs are served in the order of arrivals. Across classes (at the same station), jobs are served under a preemptive, head-of-line (HL) priority discipline, with classes 4 and 2 having higher priority over classes 1 and 3, respectively.

Similar to the notation used in §2, we denote the arrival process of class 1 (class 3) as $A_1(t)$ ($A_3(t)$), which is a (delayed) renewal process with arrival rate α_1 (α_3). The service process for class j ($j=1,\cdots,4$) is a (delayed) renewal process $S_j(t)$, which equals the number of service completions if the class has attained an amount of service time t. Let μ_j be the service rate and $m_j = 1/\mu_j$ be the mean service time for class j. The queue-length process $Q_j(t)$ ($j=1,\cdots,4$) is the number of class-j jobs in the system at time t.

Let k index a sequence of KS-RS networks, which have the same mean service times (hence with their index k omitted) but differ in their arrival rates α_j^k ($j=1,3$). Assume the arrival rates converge in the following fashion: for some α_j and θ_j ($j=1,3$),

$$k(\alpha_j^k - \alpha_j) := \theta_j^k \to \theta_j, \quad \text{as } k \to \infty. \tag{108}$$

The heavy traffic condition stipulates that both stations are bottlenecks asymptotically, i.e., the limiting traffic loads are equal to one:

$$\alpha_1 m_1 + \alpha_3 m_4 = 1 \quad \text{and} \quad \alpha_1 m_2 + \alpha_3 m_3 = 1. \tag{109}$$

Assume the above heavy traffic condition is in force for the rest of this section. Other conditions on the interarrival and service times similar to those in (10, 13, 14, 15) are also assumed.

Following Chen and Zhang [10], we have the following queue-length dynamics:

$$\begin{aligned}
Q_1^k(t) &= Q_1^k(0) + A_1^k(t) - S_1^k(T_1^k(t)),\\
Q_2^k(t) &= Q_2^k(0) + S_1^k(T_1^k(t)) - S_2^k(T_2^k(t)),\\
Q_3^k(t) &= Q_3^k(0) + A_3^k(t) - S_3^k(T_3^k(t)),\\
Q_4^k(t) &= Q_4^k(0) + S_3^k(T_3^k(t)) - S_4^k(T_4^k(t)),
\end{aligned}$$

where

$$\begin{aligned}
T_1^k(t) &= \int_0^t 1_{\{Q_4^k(s)=0, Q_1^k(s)>0\}} ds,\\
T_2^k(t) &= \int_0^t 1_{\{Q_2^k(s)>0\}} ds,\\
T_3^k(t) &= \int_0^t 1_{\{Q_2^k(s)=0, Q_3^k(s)>0\}} ds,\\
T_4^k(t) &= \int_0^t 1_{\{Q_4^k(s)>0\}} ds.
\end{aligned}$$

To describe the diffusion limit, it is convenient to define the followings,

$$\begin{aligned}
Y_1^k(t) &= 1 - T_1^k(t) - T_4^k(t) = \int_0^t 1_{\{Q_4^k(s)=0, Q_1^k(s)=0\}} ds,\\
Y_2^k(t) &= 1 - T_2^k(t) = \int_0^t 1_{\{Q_2^k(s)=0\}} ds,\\
Y_3^k(t) &= 1 - T_2^k(t) - T_3^k(t) = \int_0^t 1_{\{Q_2^k(s)=0, Q_3^k(s)=0\}} ds,\\
Y_4^k(t) &= 1 - T_4^k(t) = \int_0^t 1_{\{Q_4^k(s)=0\}} ds.
\end{aligned}$$

The strong Markov process representing the network state, $\Xi^k(t)$, is again defined by appending the residual arrival and service times to the queue-length process. It will be useful to keep in mind that here the queue-length process $Q^k(t)$ is the counterpart of the workload $W^k(t)$ in previous sections, whereas the process $Y^k(t)$ plays essentially the same regulator role as before.

Applying diffusion scaling and the usual centering to the arrival and service processes:

$$\left(\hat{\Xi}^k(t), \hat{Q}_j^k(t), \hat{A}_j^k(t), \hat{S}_i^k(t)\right) = \frac{1}{k}\left(\Xi^k(k^2t), Q_j^k(k^2t), A_j^k(k^2t) - \alpha_j^k k^2 t, S_i^k(k^2t) - \mu_i^k k^2 t\right),$$

we can rewrite the dynamics above as

$$\hat{Q}_1^k(t) = \hat{Q}_1^k(0) + \hat{X}_1^k(t) + k\alpha_1 t - k\mu_1 \tilde{T}_1^k(t) = \hat{Q}_1^k(0) + \hat{X}_1^k(t) + k\alpha_1 t + \mu_1 \hat{Y}_1^k(t) - \mu_1 \hat{Y}_4^k(t), \tag{110}$$

$$\hat{Q}_2^k(t) = \hat{Q}_2^k(0) + \hat{X}_2^k(t) + k\mu_1 \tilde{T}_1^k(t) - k\mu_2 \tilde{T}_2^k(t) \tag{111}$$

$$\begin{aligned} &= \hat{Q}_2^k(0) + \hat{X}_2^k(t) - k\mu_2 t - \mu_1 \hat{Y}_1^k(t) + \mu_2 \hat{Y}_2^k(t) + \mu_1 \hat{Y}_4^k(t), \\ \hat{Q}_3^k(t) &= \hat{Q}_3^k(0) + \hat{X}_3^k(t) + k\alpha_3 t - k\mu_3 \tilde{T}_3^k(t) = \hat{Q}_3^k(0) + \hat{X}_3^k(t) + k\alpha_3 t - \mu_3 \hat{Y}_2^k(t) + \mu_3 \hat{Y}_3^k(t), \quad (112) \\ \hat{Q}_4^k(t) &= \hat{Q}_4^k(0) + \hat{X}_4^k(t) + k\mu_3 \tilde{T}_3^k(t) - k\mu_4 \tilde{T}_4^k(t) \quad (113) \\ &= \hat{Q}_4^k(0) + \hat{X}_4^k(t) - k\mu_4 t + \mu_3 \hat{Y}_2^k(t) - \mu_3 \hat{Y}_3^k(t) + \mu_4 \hat{Y}_4^k(t), \end{aligned}$$

where

$$\begin{aligned} &\hat{X}_1^k(t) = \hat{A}_1^k(t) - \hat{S}_1^k(\tilde{T}_1^k(t)) + k(\alpha_1^k - \alpha_1)t, \\ &\hat{X}_2^k(t) = \hat{S}_1^k(\tilde{T}_1^k(t)) - \hat{S}_2^k(\tilde{T}_2^k(t)), \\ &\hat{X}_3^k(t) = \hat{A}_3^k(t) - \hat{S}_3^k(\tilde{T}_3^k(t)) + k(\alpha_3^k - \alpha_3)t, \\ &\hat{X}_4^k(t) = \hat{S}_3^k(\tilde{T}_3^k(t)) - \hat{S}_4^k(\tilde{T}_4^k(t)), \\ &\tilde{T}_j^k(t) = T_j^k(k^2 t)/k^2, \quad j = 1, \cdots, 4, \\ &\hat{Y}_j^k(t) = Y_j^k(k^2 t)/k, \quad j = 1, \cdots, 4. \end{aligned}$$

The diffusion limit for the KS-RS network is established in Chen and Zhang [10]: Assume the weak convergence $\hat{\Xi}^k(0) \Rightarrow (\hat{Q}(0), 0, 0)$ with $\hat{Q}_2(0) = \hat{Q}_4(0) = 0$, i.e., the class-2 and class-4 queue lengths and the residual arrival and service times are asymptotically zero. Then, the weak convergence $\hat{Q}^k(t) \Rightarrow \hat{Q}(t)$ holds when $k \to \infty$, if and only if the following "virtual-station condition ([16])" holds:

$$\alpha_1 m_2 + \alpha_3 m_4 < 1. \quad (114)$$

The diffusion limit $\hat{Q}(t)$ is then characterized by the following:

$$\begin{aligned} &\hat{Q}_2(t) = \hat{Q}_4(t) = 0, \quad (115) \\ &\begin{pmatrix} \hat{Q}_1(t) \\ \hat{Q}_3(t) \end{pmatrix} = \begin{pmatrix} \hat{Q}_1(0) \\ \hat{Q}_3(0) \end{pmatrix} + \begin{pmatrix} \hat{X}_1(t) \\ \hat{X}_3(t) \end{pmatrix} + R \begin{pmatrix} m_4 \hat{X}_4(t) \\ m_2 \hat{X}_2(t) \end{pmatrix} + R \begin{pmatrix} \hat{Y}_1(t) \\ \hat{Y}_3(t) \end{pmatrix} \geq 0, \quad (116) \\ &\hat{Y}_j(t) \text{ is nondecreasing in } t, \text{ with} \hat{Y}_j(0) = 0, \quad j = 1, 3, \quad (117) \\ &\int_0^\infty \hat{Q}_j(t) d\hat{Y}_j(t) = 0, \quad j = 1, 3, \quad (118) \end{aligned}$$

where

$$\begin{aligned} &\hat{X}_j(t) = \hat{A}_j(t) - \hat{S}_j(\alpha_j m_j t) + \theta_j, \quad j = 1, 3, \\ &\hat{X}_j(t) = \hat{S}_{j-1}(\alpha_{j-1} m_{j-1} t) - \hat{S}_j(\alpha_{j-1} m_j t), \quad j = 2, 4, \\ &R = \frac{1}{m_1 m_3 - m_2 m_4} \begin{pmatrix} m_3 & -m_4 \\ -m_2 & m_1 \end{pmatrix}, \end{aligned}$$

and $\hat{A}_i^k(t)$ and $\hat{S}_j^k(t)$ ($i = 1, 3$; $j = 1, \cdots, 4$) are independent (driftless) Brownian motions. (The reflection matrix R is denoted as H in Chen and Zhang [10]. Also note that in comparison with Theorem 2.1 of [10], we have re-arranged the terms involving the free processes $\hat{X}_j(t)$ slightly to facilitate discussions below.)

Similar to the remark following Theorem 1, it is important to note here that the relevant key results leading to the above diffusion limit will also play an important role below in establishing the other three sides of Figure 1 for the KS-RS network. These include: the complementarity, which holds automatically in (118), and will be used in the standard oscillation inequality shortly; the uniform attraction property (refer to the remark at the end of this subsection). Also note, the above diffusion limit for the KS-RS network parallels Theorem 1; and similarly, to support the study of the other three sides in Figure 1, variations to the diffusion limit must be established. This involves adapting Propositions 2 and 3 to the KS-RS network.

Remark. As the uniform attraction property for the KS-RS network is not given explicitly in literature, a brief account of it is in order. First, the fluid model associated with the property is as follows:

$$\begin{aligned}
&\bar{Q}_i(t) = \bar{Q}_i(0) + \alpha_i t - \mu_i \bar{T}_i(t), \quad i = 1,3,\\
&\bar{Q}_2(t) = \bar{Q}_2(0) + \mu_1 \bar{T}_1(t) - \mu_2 \bar{T}_2(t),\\
&\bar{Q}_4(t) = \bar{Q}_4(0) + \mu_3 \bar{T}_3(t) - \mu_4 \bar{T}_4(t),\\
&\bar{T}_i(0) = 0,\ \dot{\bar{T}}_i(t) \ge 0, \quad i = 1,2,3,4,\\
&\dot{\bar{T}}_1(t) + \dot{\bar{T}}_4(t) \le 1, \quad \dot{\bar{T}}_2(t) + \dot{\bar{T}}_3(t) \le 1,\\
&\dot{\bar{T}}_i(t) = 1 \ \text{ if } \bar{Q}_i(t) > 0, \quad i = 2,4,\\
&\dot{\bar{T}}_1(t) + \dot{\bar{T}}_4(t) = 1 \ \text{ if } \bar{Q}_1(t) > 0,\\
&\dot{\bar{T}}_3(t) + \dot{\bar{T}}_2(t) = 1 \ \text{ if } \bar{Q}_3(t) > 0.
\end{aligned}$$

Intuitively, the above is the deterministic counterpart of the critically loaded KS-RS network (formally, with $\alpha_1^k m_1 + \alpha_3^k m_4 = 1$ and $\alpha_1^k m_2 + \alpha_3^k m_3 = 1$ for a pre-limit network). The above also parallels the equations in (129-132) for the resource-sharing network in Appendix A, and can be derived following the standard approach, for instance, in Chen and Zhang [11] and Dai and Vande Vate [16].

Then, from (109) and (114), we note that $m_1 > m_2$ (or, $\mu_1 < \mu_2$), i.e., class-2 fluid drains away faster than the inflow from class-1. The same observation applies to class-3 and class-4 too. Hence, it is straightforward to establish the uniform attraction for the KS-RS network: there exists a time $\tau \ge 0$ such that for any solution to the the above fluid model with $|\bar{Q}(0)| \le 1$, we have $(\bar{Q}_2(t), \bar{Q}_4(t)) = 0$ and $(\bar{Q}_1(t), \bar{Q}_3(t)) = (\bar{Q}_1(\tau), \bar{Q}_3(\tau))$ for all $t \ge \tau$. Note that this version of uniform attraction is stronger than the one for the resource-sharing network in (139), and is said to be the stability of higher priority classes (SHP) in Chen and Ye [9].

With the above (strong) uniform attraction established, to justify the interchange of limits (edge IV), one can use the approach in Gurvich [26], which requires a high-moment condition on the interarrival and service times; alternatively, our approach is to verify the bounded workload condition, which we will do below. □

6.2. Stationary Distributions The deterministic DCP corresponding to the diffusion limit in (116-118) is given by:

$$\begin{pmatrix}\hat{q}_1(t)\\ \hat{q}_3(t)\end{pmatrix} = \begin{pmatrix}\hat{q}_1(0)\\ \hat{q}_3(0)\end{pmatrix} + \begin{pmatrix}\theta_1\\ \theta_3\end{pmatrix} t + R \begin{pmatrix}\hat{y}_1(t)\\ \hat{y}_3(t)\end{pmatrix}, \tag{119}$$

$$\hat{y}_j(t) \text{ is nondecreasing in } t, \text{ with } \hat{y}_j(0) = 0, \quad j = 1,3, \tag{120}$$

$$\int_0^\infty \hat{q}_j(t)\, d\hat{y}_j(t) = 0, \quad j = 1,3. \tag{121}$$

Clearly, Theorem 4 (regarding side III in Figure 1) is valid for the KS-RS network too. Indeed, since the diffusion limit and the associated deterministic DCP for the KS-RS network are all in the standard form in the sense of Dupuis and Williams [19], Theorem 2.6 there applies directly.

Theorem 7 for KS-RS network (side II) can be established in the same way via uniform stability. The key step is to prove Lemma 6, and this can be done by basically repeating the proof of the heavy traffic theorem for the KS-RS network, which requires properties such as the uniform attraction property used to prove the (conventional) diffusion limit in (115-118) for the KS-RS network. Thus, in summary, the implications in Figure 2 can be extended to a more general context (except the part in parentheses, which is specific to the resource-sharing model studied in previous sections).

6.3. Verifying Bounded Workload Condition for Interchange of Limits To establish side IV, via Theorem 14 for the KS-RS network, the main task is to verify the bounded workload condition: for some constant κ and for any $t \geq 0$,

$$\sup_{0\leq s\leq t} |\hat{Q}^k(s)| \leq \kappa \left(|\hat{Q}^k(0)| + \sup_{0\leq s\leq t} |\hat{X}^k(s)| \right). \tag{122}$$

To this end, consider any given k and $t > 0$. Let τ be the last time, before t, such that the class-2 queue is empty, i.e.,

$$\tau = \sup\{s \geq 0 : \hat{Q}_2^k(s) = 0,\ s \leq t\}.$$

Let $\tau = 0$ if the above set is empty, and denote $\hat{Q}^k(0-) = \hat{Q}^k(0)$. From (111), we have

$$\hat{Q}_2^k(t) = \hat{Q}_2^k(\tau-) + [\hat{X}_2^k(t) - \hat{X}_2^k(\tau-)] + k[\mu_1(\tilde{T}_1^k(t) - \tilde{T}_1^k(\tau-)) - \mu_2(\tilde{T}_2^k(t) - \tilde{T}_2^k(\tau-))]. \tag{123}$$

By the definition of τ, we have,

$$\tilde{T}_2^k(t) - \tilde{T}_2^k(\tau-) = t - \tau,$$

since station 2 is busy with class-2 jobs during the period $[k^2\tau, k^2t]$ (under the original time scale). From (109) and (114), we note that $m_1 > m_2$ (or, $\mu_1 < \mu_2$). Then, the second term on the right hand side of (123) is non-positive:

$$k[\mu_1(\tilde{T}_1^k(t) - \tilde{T}_1^k(\tau-)) - \mu_2(\tilde{T}_2^k(t) - \tilde{T}_2^k(\tau-))] \leq k[\mu_1(t-\tau) - \mu_2(t-\tau)] \leq 0,$$

which implies,

$$\hat{Q}_2^k(t) \leq \hat{Q}_2^k(\tau-) + \hat{X}_2^k(t) - \hat{X}_2^k(\tau-) \leq 2\left(\sup_{0\leq s\leq t} |\hat{Q}_2^k(0) + \hat{X}_2^k(s)| \right). \tag{124}$$

Similarly, we have,

$$\hat{Q}_4^k(t) \leq 2\left(\sup_{0\leq s\leq t} |\hat{Q}_4^k(0) + \hat{X}_4^k(s)| \right). \tag{125}$$

Next, we use the relationship in (111) and (113) to remove $\hat{Y}_2^k(t)$ and $\hat{Y}_4^k(t)$ in (110) and (112), and write $\hat{Q}_1^k(t)$ and $\hat{Q}_3^k(t)$ in terms of the following DCP:

$$\begin{aligned}
&\begin{pmatrix}\hat{Q}_1^k(t)\\ \hat{Q}_3^k(t)\end{pmatrix} = \begin{pmatrix}\hat{Q}_1^k(0)\\ \hat{Q}_3^k(0)\end{pmatrix} + R\begin{pmatrix}\hat{Q}_4^k(0)\\ \hat{Q}_2^k(0)\end{pmatrix} + \hat{\Sigma}^k(t) + R\cdot\begin{pmatrix}\hat{Y}_1^k(t)\\ \hat{Y}_3^k(t)\end{pmatrix} \geq 0,\\
&\quad \text{with } \hat{\Sigma}^k(t) = \begin{pmatrix}\hat{X}_1^k(t)\\ \hat{X}_3^k(t)\end{pmatrix} + R\cdot\begin{pmatrix}m_4(\hat{X}_4^k(t)-\hat{Q}_4^k(t))\\ m_2(\hat{X}_2^k(t)-\hat{Q}_2^k(t))\end{pmatrix},\\
&\hat{Y}_j^k(t) \text{ is nondecreasing in } t, \text{ with}\hat{Y}_j^k(0)=0, \quad j=1,3,\\
&\int_0^\infty \hat{Q}_j^k(t)d\hat{Y}_j^k(t) = 0, \quad j=1,3.
\end{aligned}$$

According to the standard oscillation inequality in Theorem 5.1 of [53], we have,

$$\left|\begin{pmatrix}\hat{Q}_1^k(t)\\ \hat{Q}_3^k(t)\end{pmatrix}\right| \leq \left|\begin{pmatrix}\hat{Q}_1^k(0)\\ \hat{Q}_3^k(0)\end{pmatrix} + R\begin{pmatrix}\hat{Q}_4^k(0)\\ \hat{Q}_2^k(0)\end{pmatrix}\right| + \kappa'\left(\sup_{0\leq s\leq t}|\hat{\Sigma}^k(s)|\right), \tag{126}$$

for some κ'. Combining (124, 125, 126) yields the bounded workload condition in (122) for the KS-RS network.

After verifying the bounded workload condition, the results in §4, from Lemma 9 to Theorem 14, can all be adapted in a straightforward manner (e.g., with the "usual traffic condition" replaced by the "stability of $\hat{q}(t)$"). Moreover, investigating the stability of $\hat{q}(t)$ may lead to other more explicit, model-dependent conditions on primitives. To this end, note that R is an M-matrix, and it is invertible if and only if the virtual-station condition in (114) holds. Then, the deterministic DCP in (119-121) is stable if and only if $R^{-1}\theta < 0$, i.e.,

$$\theta_1 m_1 + \theta_3 m_4 < 0 \quad \text{and} \quad \theta_1 m_2 + \theta_3 m_3 < 0. \tag{127}$$

Putting all the above together, we have the following proposition for the KS-RS network.

Proposition 15 If in addition to the heavy traffic condition in (109), the conditions in (114, 127) are satisfied, then the following properties hold.
(a) The diffusion limit $\hat{Q}(t)$ given in (115-118) is positive recurrent and has a unique stationary distribution.
(b) For any sufficiently large k, the state process $\hat{\Xi}^k(t)$ is positive recurrent and has a unique stationary distribution. Furthermore, both the state process and the stationary queue length have a finite $(p-1)$-moment: for some constant κ and for sufficiently large k,

$$\mathsf{E}|\hat{\Xi}^k(\infty)|^{p-1} \leq \kappa \quad \text{and} \quad \mathsf{E}|\hat{Q}^k(\infty)|^{p-1} \leq \kappa,$$

where the random variables, $\hat{\Xi}^k(\infty)$ and $\hat{Q}^k(\infty)$, follows the stationary distributions of $\hat{\Xi}^k(t)$ and $\hat{Q}^k(t)$, respectively.
(c) The following weak convergence of stationary distributions hold,

$$\hat{Q}^k(\infty) \Rightarrow \hat{Q}(\infty), \quad \text{as } k\to\infty,$$

where $\hat{Q}(\infty)$ follows the stationary distribution of the diffusion limit $\hat{Q}(t)$. Furthermore, for any $m\in[0,p-1)$,

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

To close this section, we remark that the sufficient condition in (106) for bounded workload applies to general multiclass queueing networks as well (e.g., those studied in Bramson [4] and Williams [54], which include the KS-RS network). Recall the workload representation from Williams ([54], equations (60, 68, 75)), for a multiclass queueing network with J stations and K classes:

$$\begin{aligned}
&\hat{W}^r(t) = \hat{W}^r(0) + R\hat{\xi}^r(t) + RCMQ\tilde{P}(\hat{\epsilon}^r(0) - \hat{\epsilon}^r(t)) + R\hat{Y}^r(t),\\
&\hat{Y}_j^r(t) \text{ is non-decreasing, with } \hat{Y}_j^r(0) = 0,\\
&\int_0^\infty \hat{W}_j^r(t) d\hat{Y}_j^r(t).
\end{aligned}$$

Here, R is the reflection matrix and is completely-S; $\hat{\xi}^r(t)$ is the "free process"; $\hat{\epsilon}^r(t) = \hat{Z}^r(t) - \Delta\hat{W}^r(t)$ is the distance from the fixed-point state space to the queue-length state $\hat{Z}^r(t)$, where the "lifting" matrix Δ maps the J-dimensional workload to the K-dimensional queue length.

The sufficient condition in (106) takes the following form, for some constant κ',

$$|\hat{\epsilon}^r(t)| \le \kappa' \left(|\hat{\epsilon}^r(0)| + \sup_{0\le s\le t} |\hat{\xi}^r(s)| \right), \quad t \ge 0.$$

Then, we can invoke the standard oscillation inequality to establish the bounded workload condition:

$$\begin{aligned}
\sup_{0\le s\le t} |\hat{W}^r(s)| &\le |\hat{W}^r(0)| + \kappa_1 \mathsf{Osc}\left(R\hat{\xi}^r(s) + RCMQ\tilde{P}\hat{\epsilon}^r(s), 0 \le s \le t \right)\\
&\le |\hat{W}^r(0)| + \kappa_2 \left(\sup_{0\le s\le t} |\hat{\xi}^r(s)| + \sup_{0\le s\le t} |\hat{\epsilon}^r(s)| \right)\\
&\le |\hat{W}^r(0)| + \kappa_2 \left(\sup_{0\le s\le t} |\hat{\xi}^r(s)| + \kappa' \left(|\hat{\epsilon}^r(0)| + \sup_{0\le s\le t} |\hat{\xi}^r(s)| \right) \right)\\
&\le \kappa'' \left(|\hat{W}^r(0)| + \sup_{0\le s\le t} |\hat{\xi}^r(s)| \right).
\end{aligned}$$

7. Concluding Remarks

As mentioned in the beginning of the paper, a main objective out this study is to establish the three sides, II, III and IV, of the rectangle in Figure 1, for a resource-sharing network under proportional fair allocation. Specifically, we have established the stationary distributions for both the diffusion limit of the workload processes, and the pre-limit processes; and we show both stationary distributions exist under the usual traffic condition. These correspond to sides II and III of the rectangle. Our approach is to establish first the *stability* of $\hat{w}(t)$, the deterministic counterpart of the diffusion limit, and then the *uniform stability* of $\hat{w}^k(t)$, the deterministic counterpart of the pre-limit networks. These stability properties are then connected to the stationary distributions of the diffusion limit and pre-limit networks. What is interesting is that the stability of $\hat{w}(t)$ implies the uniform stability of $\hat{w}^k(t)$, and hence, it suffices to establish the former, which, we show, is equivalent to the usual traffic condition.

For side IV of the rectangle, we have identified a bounded workload condition, under which the uniform stability can be strengthened to uniform p-th moment stability. The latter is sufficient for the stationary distribution of the pre-limit network to converge to the stationary distribution of the diffusion limit, thus justifying the interchange of the two limits.

These three steps (corresponding to the three sides, II, III and IV) constitute a streamlined and systematic approach to developing diffusion approximations (for stationary distributions), which has the potential to extend to a broader range of multi-class stochastic networks. To illustrate what's perhaps a first move in this direction, we have shown how to apply this approach to justify the diffusion approximation for the KS-RS network.

Appendix A: More Preliminaries As mentioned in the introduction, we collect in this Appendix A additional preliminary or technical results that supplement those in §2. These include the fluid limit and its associated uniform attraction property, the complementarity/reflection property, and the oscillation inequality associated with the DCP. We also establish a representation theorem for a tight sequence of distributions.

Results here are used in most sections of the main text and cross-referenced in the subsequent appendices, B∼D, each of which serves *exclusively* one of the main sections, §2∼§4.

A.1. Fluid Limit and Uniform Attraction We apply the standard fluid scaling to the primitive processes associated with the sequence of resource-sharing networks introduced in §2:

$$\left(\bar{E}^k(t), \bar{S}^k(t)\right) = \frac{1}{k}\left(E^k(kt), S^k(kt)\right);$$

and similarly define the fluid-scaled version of the derived processes:

$$\left(\bar{\Xi}^k(t), \bar{N}^k(t), \bar{D}^k(t), \bar{W}^k(t)\right) = \frac{1}{k}\left(\Xi^k(kt), N^k(kt), D^k(kt), W^k(kt)\right).$$

We know (e.g., [8]), with probability one,

$$\left(\bar{E}^{0,k}(t), \bar{S}^{0,k}(t)\right) \to (\lambda t, \mu t), \qquad \text{u.o.c.},$$

where $\lambda = (\lambda_r)_{r\in\mathcal{R}}$ and $\mu = (\nu_r^{-1})_{r\in\mathcal{R}}$, and the convergence is uniform on compact sets (u.o.c.) of $t \ge 0$. Observe that, for $t \ge u_r^k(1)/k$, we can write

$$\bar{E}_r^k(t) = \frac{1}{k}\left(1 + \max\left\{i - 1 : \sum_{j=2}^{i} u_r^k(j) \le k(t - u_r^k(1)/k)\right\}\right) = \frac{1}{k} + \bar{E}^{0,k}(t - u_r^k(1)/k),$$

while for $t < u_r^k(1)/k$, we have $\bar{E}_r^k(t) = 0$. Hence, we have,

$$\bar{E}_r^k(t) = \frac{1}{k}\cdot 1_{\{t \ge u_r^k(1)/k\}} + \bar{E}^{0,k}((t - u_r^k(1)/k)^+),$$

where $x^+ = \max\{x, 0\}$ for any real number x. If the residuals in initial state converge almost surely under the fluid scaling:

$$\frac{1}{k}(U^k(0), V^k(0)) \equiv \frac{1}{k}(u^k(1), v^k(1)) \to (\bar{u}(1), \bar{v}(1)) \quad \text{ as } k \to \infty,$$

then, under the assumptions (36), taking $k \to \infty$, we have

$$\left(\bar{E}^k(t), \bar{S}^k(t)\right) \to (\lambda(te - \bar{u}(1))^+, \mu(te - \bar{v}(1))^+), \qquad \text{u.o.c.} \tag{128}$$

The proposition below states that the sequence of derived processes also approaches a limit, the fluid limit. It can be shown by simply applying the "delayed" convergence in (128) to the proof of Proposition 4.2 of [57] (also Theorem 4 of [58]); hence the detailed proof is omitted.

Proposition 16 (Fluid limit) Consider the proportional fair allocation in (3). Suppose $|\bar{\Xi}^k(0)| = |\bar{W}^k(0)| + |\bar{U}^k(0)| + |\bar{V}^k(0)| \leq M$ for all k and for some constant M. Then, with probability one, for any subsequence of k, there exists a further subsequence, denoted $\mathcal{K}$, such that, along $\mathcal{K}$,

$$\bar{\Xi}_r^k(0) = \big(\bar{W}_r^k(0), \bar{U}_r^k(0), \bar{V}_r^k(0)\big) \to (\bar{w}_r(0), \bar{u}_r(1), \bar{v}_r(1)) \qquad \text{u.o.c.},$$

and

$$\big(\bar{W}^k(t), \bar{D}^k(t), \bar{N}^k(t)\big) \to \big(\bar{w}(t), \bar{d}(t), \bar{n}(t)\big) \qquad \text{u.o.c.},$$

where the (fluid) limit is Lipschitz continuous and satisfies the following: for all $r \in \mathcal{R}$, and $\ell \in \mathcal{L}$,

$$\begin{aligned} &|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq M, && (129)\\ &\bar{w}(t) = \bar{w}(0) + \rho(te - \bar{u}(1))^+ - (\bar{d}(t) - \bar{v}(1))^+ \\ &\qquad = \bar{w}(0) - \rho(te \wedge \bar{u}(1)) + (\bar{d}(t) \wedge \bar{v}(1)) + \rho t - \bar{d}(t), && (130)\\ &\bar{d}_r(t) = \int_0^t \bar{\Lambda}_r(\bar{n}(s))ds, && (131)\\ &\bar{\Lambda}_r(n) = \begin{cases} \Lambda_r(n) & \text{if } n_r > 0, \\ \rho_r & \text{if } n_r = 0. \end{cases} && (132) \end{aligned}$$

(Note that we use the convention $w_r \equiv \nu_r n_r$ and $\bar{w}_r(t) \equiv \nu_r \bar{n}_r(t)$, and repeat the definition (60) for ease of reference.)

The residuals $\bar{u}(1)$ and $\bar{v}(1)$ in the above fluid limit will cause subtle technical difficulties below. The following lemma is used to get around such difficulties. It says that after an initial transient period the residuals will vanish and have only limited impact on the workload state.

Lemma 17 Consider the fluid limit $\bar{w}(t)$ in Proposition 16. There exist a time τ and a constant κ that depend on the network parameters only (independent of k), such that

$$\begin{aligned} &(t, \bar{d}(t)) \geq (\bar{u}(1), \bar{v}(1)), \text{ for } t \geq M\tau; && (133)\\ &|\bar{w}(M\tau)| \leq \kappa M. && (134) \end{aligned}$$

Proof. Denote $\tilde{\mathcal{W}}_r = \{w : w \leq e + \rho, w_r \geq \rho_r/2\}$ for $r \in \mathcal{R}$. Since $\Lambda_r(n)$ is strictly positive and continuous on the compact set $\tilde{\mathcal{W}}_r$ (cf. Lemma 6.2(b) of Ye *et al* [57]), we have $\gamma_r^* = \inf\{\Lambda_r(n) : w \in \tilde{\mathcal{W}}_r\} > 0$. Consequently, we have $\gamma_{\min}^* = \min_{r \in \mathcal{R}} \gamma_r^* > 0$. (Again, keep in mind our convention $w_r = \nu_r n_r$ and $\bar{w}_r(t) = \nu_r \bar{n}_r(t)$.)

Let $\tau = 2 + 1/\gamma_{\min}^*$. From (129), we have $t \geq \bar{u}(1)$ for $t \geq M\tau$, the first part of (133). We claim that the inequality, $\bar{d}(t) \geq \bar{u}(1)$, also holds for $t \geq M\tau$. Otherwise, we pick any class r' such that

$$\bar{d}_{r'}(t) < \bar{u}_{r'}(1), \quad \text{for } t \leq M\tau. \tag{135}$$

Then, for any $t \in [2M, M\tau]$, we have, from (130),

$$\bar{w}_{r'}(t) = \bar{w}_{r'}(0) + \rho_{r'}(t - \bar{u}_{r'}(1))^+ \geq \rho_{r'} t \left(1 - \frac{\bar{u}_{r'}(1)}{t}\right)^+ \geq \rho_{r'} t \left(1 - \frac{\bar{u}_{r'}(1)}{2M}\right) \geq \frac{1}{2}\rho_{r'} t, \tag{136}$$

and

$$\bar{w}_r(t) \leq \bar{w}_r(0) + \rho_r(t - \bar{u}_r(1))^+ \leq t + \rho_r t, \quad \text{for all } r \in \mathcal{R}. \tag{137}$$

Hence, from (136) and (137), we have $\bar{w}(t)/t \in \tilde{\mathcal{W}}_{r'}$ for $t \in [2M, M\tau]$, which, along with the definitions of $\gamma^*_{r'}$ and $\gamma^*_{\min}$, implies $\Lambda_{r'}(\bar{n}(t)) \ge \gamma^*_{r'} \ge \gamma^*_{\min}$. Consequently, for $t \in [2M, M\tau]$, we have $\bar{\Lambda}_{r'}(\bar{n}(t)) = \Lambda_{r'}(\bar{n}(t)) \ge \gamma^*_{r'} \ge \gamma^*_{\min}$ taking into account the definition in (132) and the lower bound in (136), and finally

$$\bar{d}_{r'}(M\tau) \ge \int_{2M}^{M\tau} \bar{\Lambda}_{r'}(\bar{n}(t)) \ge (M\tau - 2M)\gamma^*_{\min} = M \ge \bar{v}_{r'}(1).$$

This contradicts (135).

Observe that the inequality in (137) holds generally (not requiring the assumption that $\bar{d}_{r'}(t) < \bar{u}_{r'}(1)$), which implies the estimation in (134). □

The next proposition, concerning the uniform attraction of the fluid limit $\bar{w}(t)$, is *almost* a paraphrase of the (more general) result established in Ye and Yao ([58], Theorem 6) specialized to the setting here; also refer to a similar result in Kelly and Williams ([38], Theorem 5.2) and Kang *et al* [32]. The only exception is that in their papers there is no residual initial arrival times and service requirements involved. However, the effect of such residuals can be mitigated easily by the lemma just established. The proof is hence omitted.

Proposition 18 Consider the fluid limit $\bar{w}(t)$ in Proposition 16 and Lemma 17, along with the constant M and τ specified there. Assume the heavy traffic condition in (11) holds.
(a) The link-based workload, $A_\ell \bar{w}(t)$ ($\ell \in \mathcal{L}$), is non-decreasing in time $t \ge M\tau$; and there exists a constant κ_w that only depends on the network parameters, such that the following bounds hold for all $t \ge 0$,

$$|\bar{w}(t)| \le \kappa_w(|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)|) \ (\le \kappa_w M). \tag{138}$$

(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:

$$|\bar{w}(t) - w^*| \le \epsilon, \quad \text{for } t \ge T_{M,\epsilon}. \tag{139}$$

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

$$d^{fp}(\bar{w}(t)) \le |G^T(\bar{w}(t) - w^*)| + |H^T \bar{w}(t)| \le \epsilon, \quad \text{for } t \ge T_{M,\epsilon}. \tag{140}$$

(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 $t \ge 0$.

A.2. Reflection near the Boundary of $\mathcal{W}$ 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 (26).

Lemma 19 (Lemma 2 of Ye and Yao [60]) 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

$$|w| \le \kappa \ \text{ and } \ d^{fp}(w) \le \sigma, \tag{141}$$

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

$$g_\ell^T w > \epsilon \qquad \Rightarrow \qquad A_\ell \Lambda(n) = \sum_{r \in \mathcal{R}} a_{\ell r} \Lambda_r(n) = c_\ell. \tag{142}$$

In words, the link ℓ will be fully occupied if the workload state of the network is away from the ℓ-facet, and toward the interior, of the fixed-point state space.

A.3. Dynamic Complementarity Problem and Oscillation Inequality What connects the workload process with its limiting regime under diffusion scaling in the diffusion limit theorems is the following DCP:

$$\begin{aligned}
&w(t) = w(0) + x(t) + BGy(t) + BHz(t) \ (\geq 0), \quad \text{for } t \geq 0; && (143)\\
&G^T w(t) \geq 0, \quad \text{for } t \geq 0; && (144)\\
&y_\ell(t) \text{ is non-decreasing in } t \geq 0,\ y_\ell(0) = 0, \quad \ell \in \mathcal{L}; && (145)\\
&\int_0^\infty w(t)^T G \ dy(t) = 0; && (146)\\
&H^T w(t) = 0, \quad \text{for } t \geq 0; && (147)\\
&z(0) = 0; && (148)
\end{aligned}$$

where the initial state $w(0)$ and the "free process" $x(t)$ (an R-dimensional function) are given. Clearly, the deterministic DCP in (39-44) is a special case of the above DCP.

Proposition 20 (Ye and Yao [60]) Given an R-dimensional RCLL "free process" $x(t)$, there exists a unique solution, $(w(t), y(t), z(t))$, to the DCP in (143)-(148).

The oscillation inequality is a useful tool to establish the boundedness of the workload process below (refer to Lemma 23(c) below). For any RCLL (vector) function $f(u)$ $(u \geq 0)$ and any time interval $[s,t]$, denote

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

Lemma 21 (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 $0 \leq s \leq t$,

$$\mathsf{Osc}(G^T w(\cdot), [s,t]) \text{ and } \mathsf{Osc}(y(\cdot), [s,t]) \leq \kappa_c(\mathsf{Osc}(x(\cdot), [s,t]) + \epsilon). \tag{149}$$

If in addition,

$$|H^T w(t)| \leq \epsilon, \quad \text{for } t \geq 0, \tag{150}$$

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

$$\mathsf{Osc}(w(\cdot), [s,t]) \text{ and } \mathsf{Osc}(y(\cdot), [s,t]) \leq \kappa_c(\mathsf{Osc}(x(\cdot), [s,t]) + \epsilon). \tag{151}$$

The first oscillation inequality in the above lemma is essentially Proposition 7.1 of [32], a variation of Theorem 5.1 of [53]. The second is a direct consequence of the first and the additional condition in (150). Hence the detailed proof is omitted.

A.4. A Representation Theorem The following lemma (used at the beginning of the proof of Proposition 2) is an independent result, which relates the tightness of a sequence of distributions to a uniform bound through a coupling technique — in the same spirit as the Skorohod representation theorem.

Lemma 22 A sequence (of random variables/vectors) $\{w_k\}$ is tight, if and only if the following statement holds: There exists another sequence $\{\tilde{w}_k\}$, with $\tilde{w}_k \overset{\mathrm{d}}{=} w_k$ for each k, and all $\tilde{w}_k$'s coupled on a common probability space Ω, such that for every $\omega \in \Omega$, $\{\tilde{w}_k\}$ is uniformly bounded, i.e., $\sup_k |\tilde{w}_k(\omega)| < \infty$.

Proof. The "if" part holds trivially. To prove the "only if" part, we first consider the case that w_k's are nonnegative scalars (random variables). Let $F_k(x)$ be the distribution function of w_k; and denote

$$F_k^{-1}(u) = \inf\{t \geq 0 : F_k(t) \geq u\}, \qquad u \in [0,1).$$

Let ξ be a random variable uniformly distributed over the interval $[0,1) = \Omega$. Define $\tilde{w}_k = F_k^{-1}(\xi)$; then, $\tilde{w}_k \overset{\mathrm{d}}{=} w_k$,

Now, let $\omega \in [0,1) = \Omega$ be any given sample. Tightness implies there exists a finite (but sufficiently large) $M = M(\omega)$, which may depend on ω, but independent of k, such that $F_k(M) \geq \omega$ for all k. Hence, for the given ω, we have $\tilde{w}_k(\omega) = F_k^{-1}(\omega) \leq M$ for all k; i.e., $\sup_k \tilde{w}_k(\omega) < \infty$.

In general (i.e., regardless whether w_k's are non-negative or not, scalars or vectors), let $S_k = |w_k|$ (the absolute value of the sum of w_k's components), a non-negative scalar. Then, the above applies to S_k; hence, we have $\tilde{S}_k \overset{\mathrm{d}}{=} S_k$ for every k, and $\sup_k \tilde{S}_k(\omega) < \infty$ for every $\omega \in \Omega$.

To construct $\tilde{w}_k$, first, for each k, define an independent sequence $\{\tilde{w}_{k,m}, m = 0, 1, \cdots\}$ such that $\tilde{w}_{k,m}$ follows the distribution of w_k conditioned on $\{m \leq S_k < m+1\}$, i.e., for any vector x,

$$\mathsf{P}(\tilde{w}_{k,m} \leq x) = \mathsf{P}(w_k \leq x | m \leq S_k < m+1).$$

Given the definition that $S_k = |w_k|$, the above implies

$$\mathsf{P}(m \leq |\tilde{w}_{k,m}| < m+1) = 1.$$

Moreover, by redefining the value of $\tilde{w}_{k,m}(\omega)$ (at most) on an event of zero probability, we can strengthen the above so that for all $\omega \in \Omega$,

$$m \leq |\tilde{w}_{k,m}| < m+1. \tag{152}$$

Next, let $\tilde{w}_k = \sum_{m=0}^{\infty} \tilde{w}_{k,m} \mathbf{1}\{m \leq \tilde{S}_k < m+1\}$. Then, for $m = 0, 1, \cdots$, we have

$$\mathsf{P}(\tilde{w}_k \leq x | m \leq \tilde{S}_k < m+1) = \mathsf{P}(\tilde{w}_{k,m} \leq x) = \mathsf{P}(w_k \leq x | m \leq S_k < m+1).$$

Therefore,

$$\mathsf{P}(\tilde{w}_k \leq x) = \sum_{m=0}^{\infty} \mathsf{P}(\tilde{w}_k \leq x | m \leq \tilde{S}_k < m+1) \mathsf{P}(m \leq \tilde{S}_k < m+1)$$
$$= \sum_{m=0}^{\infty} \mathsf{P}(w_k \leq x | m \leq S_k < m+1) \mathsf{P}(m \leq S_k < m+1) = \mathsf{P}(w_k \leq x).$$

That is, $\tilde{w}_k \overset{\mathrm{d}}{=} w_k$.

Now, given any $\omega \in \Omega$, choose an integer $M = M(\omega)$ such that $\sup_k \tilde{S}_k(\omega) \leq M$. Then, from its definition, $\tilde{w}_k$ can be written as $\tilde{w}_k = \sum_{m=0}^{M-1} \tilde{w}_{k,m} \mathbf{1}\{m \leq \tilde{S}_k < m+1\}$. This, along with the property in (152), implies that $\sup_k |\tilde{w}_k(\omega)| \leq M$. □

Appendix B: Proofs for §2

B.1. Proof of Proposition 2 Let's first lay out the setup before presenting the proof. From the functional central limit theorem (i.e., Donsker's theorem; refer to [1]), we know the diffusion-scaled processes, $\hat{E}^{0,k}_r(t)$ and $\hat{S}^{0,k}_r(t)$, converge weakly to Brownian motions:

$$\hat{E}^{0,k}_r(t) \Rightarrow \hat{E}_r(t) \quad \text{and} \quad \hat{S}^{0,k}_r(t) \Rightarrow \hat{S}_r(t), \tag{153}$$

where $\hat{E}_r(t)$ and $\hat{S}_r(t)$ are zero-mean Brownian motions with variances $\lambda_r^3\sigma^2_{a,r}$ and $\nu_r^{-3}\sigma^2_{s,r}$, respectively.

Following the Skorohod representation theorem, we can turn the weak convergence into a probability one convergence of suitable copies of processes (refer to [1]). Hence, we now assume that the following convergence has replaced the weak convergences in (153): with probability one, as $k \to \infty$,

$$\hat{E}^{0,k}_r(t) \to \hat{E}_r(t) \text{ and } \hat{S}^{0,k}_r(t) \to \hat{S}_r(t) \quad \text{u.o.c. of } t \geq 0. \tag{154}$$

Per Lemma 22, we can replace the tightness of initial states $\{\hat{\Xi}^k(0)\}$ by the uniform boundedness, and hence assume that the sequence $\{\hat{\Xi}^k(0)\}$ is uniformly bounded: for each sample, there exists a (sample-dependent) constant M such that

$$|\hat{\Xi}^k(0)| \leq M \quad \text{for all } k. \tag{155}$$

In addition, we can assume that all these variables and processes are coupled on a common probability space. Having invoked the representations for weak convergence and tightness, we now consider another version of the sequence of networks, in which the k-th network evolves following the same probabilistic law of the k-th network in the original version (i.e., the sequence of networks referred to in the theorem). Hence, the probabilistic property are the same for the k-th networks in the two versions of network sequences.

Observing that $t^k_0 > M/k \geq (|u^k(1)| + |v^k(1)|)/k^2$ for sufficiently large k (since $kt^k_0 \to \infty$), we can write

$$\hat{E}^k_r(t^k_0 + t) - \hat{E}^k_r(t^k_0) = \hat{E}^{0,k}_r(t^k_0 + t - u^k_r(1)/k^2) - \hat{E}^{0,k}_r(t^k_0 - u^k_r(1)/k^2).$$

Note that $t^k_0 - u^k_r(1)/k^2 \to 0$. Then, applying (154) to the right-hand-side of the above, we have

$$\hat{E}^k_r(t^k_0 + t) - \hat{E}^k_r(t^k_0) \to \hat{E}_r(t) \tag{156}$$

As a direct consequence of (155), the initial states of the fluid-scaled processes, $\tilde{\Xi}^k(t) := \Xi^k(k^2 t)/k^2$, will vanish:

$$\tilde{\Xi}^k_r(0) = \frac{1}{k}\hat{\Xi}^k(0) \to 0, \qquad \text{as } k \to \infty. \tag{157}$$

Note that Propositions 16 and 18 remain unchanged (with the residuals $\bar{u}(1)$ and $\bar{v}(1)$ set to zeros) if we replace the scaling factor k by k^2; therefore, from Proposition 18(b), we have, under the proportional fair allocation scheme,

$$\tilde{D}^k_r(t) \to \rho_r t, \quad \text{u.o.c. of } t \geq 0. \tag{158}$$

From (154) and (158) and using the similar argument for (156), we have

$$\hat{S}^k_r(\tilde{D}^k_r(t^k_0 + t)) - \hat{S}^k_r(\tilde{D}^k_r(t^k_0)) \to \hat{S}_r(\rho_r t), \quad \text{u.o.c. of } t \geq 0, \tag{159}$$

where $\hat{S}_r(\rho_r t)$ is a Brownian motion with zero mean and variance $\rho_r \nu_r^{-3}\sigma_{s,r}^2$. Consequently, $\hat{X}_r^k(t)$ converges as follows,

$$\hat{X}_r^k(t_0^k+t)-\hat{X}_r^k(t_0^k)\to\hat{X}_r(t)=\theta_r t+\nu_r\left(\hat{E}_r(t)-\hat{S}_r(\rho_r t)\right),\quad \text{u.o.c. of } t\ge 0, \tag{160}$$

with the limit $\hat{X}_r(t)$ being a Brownian motion with drift and variance coefficients, θ_r and σ_r^2, specified in (10) and (34).

Below we shall focus on a fixed sample for which the above u.o.c. convergence in (154, 156, 158-160) and the boundedness property in (155) hold.

Consider the time interval $[\tau,\tau+\delta]$, where $\tau\ge 0$ and $\delta>0$ can be chosen arbitrarily. Let $T>0$ be a fixed time of a certain magnitude to be specified later. Let the index k be a large integer. Divide the time interval $[\tau,\tau+\delta]$ into a total of $\lceil k\delta/T\rceil$ segments with equal length T/k, where $\lceil\cdot\rceil$ denotes the integer ceiling. The j-th segment, $j=0,...,\lceil k\delta/T\rceil-1$, covers the time interval $[\tau+jT/k,\tau+(j+1)T/k]$. Note that the last interval (with j= $\lceil k\delta/T\rceil-1$) covers a negligible piece of time beyond the right end of $[\tau,\tau+\delta]$ if $k\delta/T$ is not an integer. For notational simplicity, below we shall assume $k\delta/T$ to be an integer (i.e., omit the ceiling notation). Then, for any $t\in[\tau,\tau+\delta]$, we can write it as $t=\tau+(jT+u)/k$ for some $j=0,\cdots,k\delta/T$ and $u\in[0,T]$. Therefore, we write

$$\hat{W}^k(t)=\hat{W}^k(\tau+\frac{jT+u}{k})=\frac{1}{k}W^k((k^2\tau+kjT)+ku):=\bar{W}^{k,j}(u),\qquad u\in[0,T],\ j\le\frac{k\delta}{T}. \tag{161}$$

That is, for each time point t, we will study the behavior of $\hat{W}^k(t)$ through the fluid process, $\bar{W}^{k,j}(u)$, over the time interval $u\in[0,T]$. Similarly define $\bar{\Xi}^{k,j}(u)$, $\bar{U}^{k,j}(u)$, $\bar{V}^{k,j}(u)$, $\bar{N}^{k,j}(u)$ and $\bar{Y}^{k,j}(u)$ as the fluid "magnifiers" of $\hat{\Xi}^k(t)$, $\hat{U}^k(t)$, $\hat{V}^k(t)$, $\hat{N}^k(t)$ and $\hat{Y}^k(t)$. The above representation follows the idea of hydrodynamics in Bramson [4] (also refer to [42, 50, 58]).

We need the following lemma (which is a modification of Lemma 7 in [60]), with its proof deferred to the next subsection.

Lemma 23 Consider the time interval $[\tau,\tau+\delta]$, with $\tau\ge 0$ and $\delta>0$; pick a constant $C>0$ such that

$$\mathsf{Osc}(\hat{X}(\cdot),[\tau,\tau+\delta])\le C; \tag{162}$$

and suppose

$$\sup_k|\hat{\Xi}^k(\tau)|=\sup_k(|\hat{W}^k(\tau)|+|\hat{U}^k(\tau)|+|\hat{V}^k(\tau)|)\le M, \tag{163}$$

for some constant $M\ge 0$. Let $\epsilon>0$ be any given (small) number. Then, there exists a sufficiently large T such that, for sufficiently large k, the following results hold for all *positive* integers $j=1,\cdots,k\delta/T$ (excluding $j=0$):

(a) (uniform attraction)

$$d^{fp}(\bar{W}^{k,j}(u))\le\epsilon,\qquad \text{for all } u\in[0,T]; \tag{164}$$

(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)

$$|\bar{W}^{k,j}(u)|\le\kappa_w|\hat{\Xi}^k(\tau)|+\kappa_x(C+\epsilon)\le\kappa=\kappa_w M+\kappa_x(C+\epsilon),\qquad \text{for all } u\in[0,T], \tag{165}$$

where κ_w is a positive constant specified in Proposition 18, and κ_x is a positive constant that depends only on network parameters.

Proof (of Proposition 2). Keeping in mind the boundedness in (155), we know that the condition in (163) is satisfied with $\tau=0$. Hence, we can apply Lemma 23, with $\tau=0$, to the processes $\hat{\Xi}^k(t)$. From the definition of t_0^k, we have for sufficiently large k,

$$\hat{W}^k(t_0^k)=\bar{W}^{k,j}(u), \quad \text{for some } u\in[0,T], \text{ and some } j \text{ satisfying } 1\le j<k\delta/T.$$

Then, by Part (c) of the lemma, we have

$$|\hat{W}^k(t_0^k)|\le \kappa_w|\hat{\Xi}^k(0)|+\kappa_x(C+\epsilon), \tag{166}$$

and hence

$$\limsup_{k\to\infty}|\hat{W}^k(t_0^k)|\le \limsup_{k\to\infty}\kappa_w|\hat{\Xi}^k(0)|+\kappa_x(C+\epsilon).$$

Observe that the left hand side is independent of ϵ and C, and that C can be made arbitrarily small (for each given sample-path) by choosing a sufficiently small δ in the lemma. Therefore, the above inequality can be strengthened as

$$\limsup_{k\to\infty}|\hat{W}^k(t_0^k)|\le \limsup_{k\to\infty}\kappa_w|\hat{\Xi}^k(0)|\le \kappa_w M. \tag{167}$$

Since $\{\hat{W}^k(0)\}$ is tight (cf. the condition in (155)), the above inequality implies that the sequence $\{\hat{W}^k(t_0^k)\}$ is also tight. Moveover, from Lemma 23(a), we have

$$d^{fp}(\hat{W}^k(t_0^k))\to 0 \quad \text{ as } k\to\infty. \tag{168}$$

From the tightness of $\{\hat{W}^k(t_0^k)\}$ and the convergence in (168), we know for any subsequence of k, there exists a further subsequence, denoted $\mathcal{K}$, such that

$$\hat{W}^k(t_0^k)\Rightarrow \hat{W}(0)\in\mathcal{W} \quad \text{ as } k\to\infty \text{ along } \mathcal{K}. \tag{169}$$

Moreover, we observe

$$(\hat{U}^k(t_0^k),\hat{V}^k(t_0^k))\Rightarrow 0 \quad \text{ as } k\to\infty \text{ along } \mathcal{K}. \tag{170}$$

To see this, note that for any fixed class r and time point $t'>0$, we have $1<E_r^{0,k}(k^2t_0^k)<\lambda_r k^2 t'$ for sufficiently large k. Thus, the residual interarrival time $U_r^k(k^2t_0^k)$, as a portion of the interarrival time of the $E_r^{0,k}(k^2t_0^k)$-th arrival, is dominated by the maximum interarrival times of the second through the $(\lambda_r k^2 t')$-th arrivals (but excluding the first arrival); refer to the equations in (4, 5, 9). From Lemma 5.1 of [4], we know that $\hat{U}_r^k(t_0^k)=U_r^k(k^2t_0^k)/k$ approaches zero, i.e., the first convergence in (170). The convergence of $\hat{V}^k(t_0^k)$ in (170) is justified in the same manner.

Now we "restart" the k-th network from the time epoch t_0^k. Given the above initial conditions in (169) and (170) for the sequence $\{\hat{W}^k(t_0^k+t),\ k\in\mathcal{K}\}$, we can apply Theorem 1 to the subsequence $\mathcal{K}$ to conclude Proposition 2, except for the property in (38), which is shown below.

To simplify notation, denote $x_k=|\hat{W}^k(t_0^k)|$ and $y_k=\kappa_w|\hat{\Xi}^k(0)|$ in the rest of this proof. Then, the inequality in (166), which is essentially the second inequality in (165), reads

$$x_k-y_k\le \kappa_x(C+\epsilon).$$

Applying the same argument that leads to (167), we have, with probability one,

$$\limsup_{k\to\infty}(x_k-y_k)\le 0,$$

which implies for any $\sigma > 0$,

$$\mathsf{P}\left(\bigcup_{k'>0}\bigcap_{k>k'}\{x_k - y_k \le \sigma\}\right) = 1, \quad \text{or} \quad \lim_{k'\to\infty} \mathsf{P}\left(\bigcap_{k>k'}\{x_k - y_k \le \sigma\}\right) = 1.$$

Then, for any $\delta > 0$, there exists a sufficiently large index k' such that

$$\mathsf{P}\left(\bigcap_{k>k'}\{x_k - y_k \le \sigma\}\right) \ge 1 - \delta;$$

and hence, for any $k > k'$,

$$\mathsf{P}\{x_k - y_k \le \sigma\} \ge 1 - \delta.$$

Observe that for any positive constant M, the following inequality holds,

$$\mathsf{P}\{x_k \le M + \sigma\} + \mathsf{P}\{y_k > M\} \ge \mathsf{P}\{x_k - y_k \le \sigma\}.$$

Combining the last two inequalities, we have

$$\mathsf{P}\{x_k \le M + \sigma\} \ge \mathsf{P}\{y_k \le M\} - \delta.$$

Hence,

$$\limsup_{k\to\infty} \mathsf{P}\{x_k \le M + \sigma\} \ge \limsup_{k\to\infty} \mathsf{P}\{y_k \le M\} - \delta.$$

Since δ is arbitrarily chosen, the above is equivalent to

$$\limsup_{k\to\infty} \mathsf{P}\{x_k \le M + \sigma\} \ge \limsup_{k\to\infty} \mathsf{P}\{y_k \le M\}.$$

If we restrict the index k to the subsequence $\mathcal{K}$, then given the weak convergence in (37) (i.e., (169)), the above is reduced to

$$\mathsf{P}\{|\hat{W}(0)| \le M + \sigma\} \ge \limsup_{k\to\infty, k\in\mathcal{K}} \mathsf{P}\{\kappa_w|\hat{\Xi}^k(0)| \le M\}. \tag{171}$$

As σ is arbitrary, the above implies the second inequality in (38). □

B.1.1. Proof of Lemma 23 This proof is a modification of the proof of Lemma 7 in [60], taking into account that we must deal with the initial residuals here. To this end, we need a variation of Propositions 16 and 18, summarized in the lemma below. (The proof of the lemma is omitted as it parallels that of Lemma 12 of [60], using Lemma 17 above to handle the initial residuals.)

Lemma 24 Let $M > 0$ be an any given constant, and j_k an integer in $[0, k\delta/T]$ for each k. Suppose $|\bar{\Xi}^{k,j_k}(0)| \le M$ for sufficiently large k. Then, for any subsequence of k, there exists a further subsequence such that along the subsequence, the process $(\bar{W}^{k,j_k}(t), \bar{D}^{k,j_k}(t), \bar{U}^{k,j_k}(0), \bar{V}^{k,j_k}(0))$ converge u.o.c. to the fluid limit $(\bar{w}(t), \bar{d}(t), \bar{u}(1), \bar{v}(1))$ that satisfies all the properties described in Propositions 16 and 18. If in addition $j_k \ge 1$ for sufficiently large k, then $\bar{u}(1) = \bar{v}(1) = 0$.

We specify the time length of T as follows:

$$T \geq \max\{T_{\kappa,\epsilon/8}, T_{\kappa,\sigma/2}\}, \tag{172}$$

where the terms on the right hand side are defined in Proposition 18, and $\sigma = \sigma(\kappa', \epsilon/2)$ is specified in Lemma 19 (with κ and ϵ given in the current lemma under proof, and κ' being a constant that depends on network parameters only and will be specified shortly). Note that T is large enough so that in the fluid network in Proposition 18 (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 $\bar{w}(0)$ that is bounded by κ.

Step 1. We use contradiction to prove the three parts of Lemma 23, (a,b,c), for $j = 1$. Suppose, 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 $j = 1$. To reach a contradiction, below we will construct an infinite subsequence $\mathcal{K}_2 \subset \mathcal{K}_1$, such that the desired properties in (a,b,c) hold for $j = 1$ for sufficiently large $k \in \mathcal{K}_2$.

By Lemma 24 (and Proposition 16), there exists a further subsequence $\mathcal{K}_2 \subset \mathcal{K}_1$ such that, as $k \to \infty$ along $\mathcal{K}_2$,

$$\bar{W}^{k,0}(u) \to \bar{w}(u) \qquad \text{u.o.c. of } u \geq 0, \tag{173}$$

and

$$(\bar{U}^{k,0}(0), \bar{V}^{k,0}(0)) \to (\bar{u}(1), \bar{v}(1)),$$

with $|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)| \leq M$ $(\leq \kappa)$. Since $T \geq T_{\kappa,\epsilon/8}$, applying the uniform attraction property in Lemma 24 (and Proposition 18) to the above limit yields:

$$d^{fp}(\bar{w}(u)) \leq \frac{\epsilon}{8} \ \text{ for all } u \geq T; \ \text{ and } \ |\bar{w}(u)| \leq \kappa_w(|\bar{w}(0)| + |\bar{u}(1)| + |\bar{v}(1)|) \ \text{ for all } u \geq 0. \tag{174}$$

Note that $\bar{W}^{k,0}(T+u) \equiv \bar{W}^{k,1}(u)$ (and $\bar{W}^{k,0}(0) = \hat{W}^k(\tau)$). Hence, the convergence in (173), along with (174), implies that (a,c) holds with $j = 1$ for sufficiently large $k \in \mathcal{K}_2$. Specifically, the bounding property in (c) is reduced to the following, for $j = 1$,

$$|\bar{W}^{k,1}(u)| = |\bar{W}^{k,0}(T+u)| \leq \kappa_w |\hat{\Xi}^k(\tau)| + \epsilon, \qquad \text{for all } u \in [0, T]. \tag{175}$$

As a by-product (used below), we have for $u \in [0, T]$,

$$|\bar{W}^{k,1}(u)| \ \leq \ \kappa_w M + \epsilon \leq \kappa_w \kappa + \epsilon := \kappa'. \tag{176}$$

Furthermore, since $T \geq T_{\kappa,\sigma/2}$, the first inequality in (174) 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 $j = 1$ for sufficiently large $k \in \mathcal{K}_2$, i.e.,

$$d^{fp}(\bar{W}^{k,1}(u)) \leq \sigma, \quad \text{for } u \in [0, T]. \tag{177}$$

In addition to (174), we can require the following via Lemma 24 (and Proposition 18) too:

$$|G^T(\bar{w}(T+u) - w^*)| \leq \frac{\epsilon}{8}, \tag{178}$$

$$|G^T(\bar{W}^{k,1}(u) - \bar{w}(T+u))| = |G^T(\bar{W}^{k,0}(T+u) - \bar{w}(T+u))| \leq \frac{\epsilon}{8}. \tag{179}$$

for $u \in [0,T]$ and for some fixed-point state w^*. Now, consider any link ℓ satisfying the "if" condition in (b) for $j=1$. Using the estimations in (178) and (179), we have for sufficiently large $k (\in \mathcal{K}_2)$ and for any $u \in [0,T]$,

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

and hence

$$g_\ell^T \bar{W}^{k,1}(u) \ge g_\ell^T \bar{N}^{k,1}(u') - \frac{\epsilon}{2} \ge \frac{\epsilon}{2}. \tag{180}$$

Thereafter, we have

$$\bar{Y}_\ell^{k,1}(u) - \bar{Y}_\ell^{k,1}(0) = \int_0^u \left(c_\ell - A_\ell \Lambda(\bar{W}^{k,1}(s)) \right) ds = 0, \tag{181}$$

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 19 to the link ℓ given the upper bound in (176) and the estimations in (177) and (180).

Step 2. We now extend the above to $j = 2, \ldots, k\delta/T$. Suppose again, to the contrary, there exists a subsequence $\mathcal{K}_1$ of k such that, for any $k \in \mathcal{K}_1$, at least one of the results in (a,b,c) does not hold for some integer $j \in [2, k\delta/T]$. Consequently, for any $k \in \mathcal{K}_1$, there exists a smallest positive integer j_k in the interval $[2, k\delta/T]$ such that at least one of the properties in (a, b, c) does not hold. To reach a contradiction, in the rest of the proof we will construct an infinite subsequence $\mathcal{K}_2 \subset \mathcal{K}_1$, such that the desired properties in (a, b, c) hold for $j = j_k$ for sufficiently large $k \in \mathcal{K}_2$.

Following the earlier argument, under the (contradictory) assumption above, the results in (a,b,c) hold for $j = 1, ..., j_k - 1$, for each $k \in \mathcal{K}_1$. Specifically, for $j = j_k - 1$ (≥ 1), we have

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

By Lemma 24 (and Proposition 16), there exists a further subsequence $\mathcal{K}_2 \subset \mathcal{K}_1$ such that

$$\bar{W}^{k,j_k-1}(u) \to \bar{w}(u) \qquad \text{u.o.c. of } u \ge 0, \text{ as } k \to \infty \text{ along } \mathcal{K}_2, \tag{182}$$

with $|\bar{w}(0)| \le \kappa$. Since $T \ge T_{\kappa, \epsilon/8}$, applying the uniform attraction property in Lemma 24 (and Proposition 18) to the above limit yields:

$$d^{fp}(\bar{w}(u)) \le \frac{\epsilon}{8} \ \text{ for all } u \ge T. \tag{183}$$

Note that $\bar{W}^{k,j_k-1}(T+u) \equiv \bar{W}^{k,j_k}(u)$. Hence, the convergence in (182), along with (183), implies that (a) holds with $j = j_k$ for sufficiently large $k \in \mathcal{K}_2$.

In addition to (183), we can claim the following via Lemma 24 (and Proposition 18) too:

$$|\bar{w}(u)| \le \kappa_w |\bar{w}(0)| \le \kappa_w \kappa, \ \text{ for all } u \ge 0; \tag{184}$$

$$|G^T(\bar{w}(T+u) - w^*)| \le \frac{\epsilon}{8}, \ \text{ for } u \ge 0, \text{ and for some } w^* \in \mathcal{W}; \tag{185}$$

$$\begin{aligned} &|G^T(\bar{W}^{k,j_k}(u) - \bar{w}(T+u))| = |G^T(\bar{W}^{k,j_k-1}(T+u) - \bar{w}(T+u))| \le \frac{\epsilon}{8}, \\ &\qquad \text{for } u \in [0,T], \text{ and for sufficiently large } k \in \mathcal{K}_2. \end{aligned} \tag{186}$$

The first bound above implies the following for sufficiently large $k\in\mathcal{K}_2$ and for all $u\in[0,T]$:

$$|\bar{W}^{k,j_k}(u)|\le|\bar{w}(T+u)|+\epsilon\le\kappa_w|\bar{w}(0)|+\epsilon\le\kappa_w\kappa+\epsilon=\kappa'. \tag{187}$$

Furthermore, since $T\ge T_{\kappa,\sigma/2}$, the inequality in (183) 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 $j=j_k$ for sufficiently large $k\in\mathcal{K}_2$, i.e.,

$$d^{fp}(\bar{W}^{k,j_k}(u))\le\sigma,\quad\text{for } u\in[0,T]. \tag{188}$$

Now, consider any link ℓ satisfying the "if" condition in (b) for $j=j_k$. Similar to (180) and (181), we use the estimations in (185) and (186) to show that for sufficiently large $k(\in\mathcal{K}_2)$ and for any $u\in[0,T]$,

$$g_\ell^T\bar{W}^{k,j_k}(u)\ge\bar{W}^{k,j_k}(u')-\frac{\epsilon}{2}\ge\frac{\epsilon}{2}, \tag{189}$$

and thereafter apply the estimations in (187, 188, 189) to derive the following,

$$\bar{Y}_\ell^{k,j_k}(u)-\bar{Y}_\ell^{k,j_k}(0)=\int_0^u\big(c_\ell-A_\ell\bar{\Lambda}(\bar{N}^{k,j_k}(s))\big)\,ds=0. \tag{190}$$

Consider any sufficiently large $k\in\mathcal{K}_2$, such that the results in (a) and (b) hold for $j=1,\cdots,j_k$ (but (a) may not holds for $j=0$). This implies that the processes, $(w(t),x(t),y(t),z(t))=(\hat{W}^k(t),\hat{X}^k(t),\hat{Y}^k(t),\hat{Z}^k(t))$, satisfy the specifications in Lemma 21 for $t\in[\tau+T/k,\tau+(j_kT+T)/k]$. Hence, we have for any $t\in[\tau+T/k,\tau+(j_kT+T)/k]$ $(\subset[\tau,\tau+\delta])$,

$$\mathsf{Osc}(\hat{W}^k(\cdot),[\tau+T/k,t])\le\kappa_c(\mathsf{Osc}(\hat{X}^k(\cdot),[\tau+T/k,t])+\epsilon)=\kappa_c(C+\epsilon). \tag{191}$$

Consequently, we have the following estimations,

$$|\hat{W}^k(t)|\le|\hat{W}^k(\tau+T/k)|+\mathsf{Osc}(\hat{W}^k(\cdot),[\tau+T/k,t])\le\kappa_w|\hat{W}^k(\tau)|+\epsilon+\kappa_c(C+\epsilon),$$

where in the second inequality we have also applied the conclusion in (175), i.e., $|\hat{W}^k(\tau+T/k)|=|\bar{W}^{k,1}(0)|\le\kappa_w|\hat{W}^k(\tau)|+\epsilon$. Keeping in mind that $\bar{W}^{k,j_k}(u)\equiv\hat{W}^k(\tau+(j_kT+u)/k)$ and with $\kappa_x=\kappa_c+1$, the above implies that (c) holds with $j=j_k$ for sufficiently large $k\in\mathcal{K}_2$. □

B.2. Proof of Proposition 3 In this proof, we use the superscript m_k to denote the mixed scaling. For example, we write for the k-th network:

$$\hat{W}^{m_k}(t)=\frac{1}{m_k}\hat{W}^k(m_kt).$$

Other processes are similarly indexed.

The proof of Proposition 3 is a straightforward modification of the arguments that establish the diffusion limits in Theorem 1 and Proposition 2. In particular, all hat-processes in the proof are now indexed by m_k. Below, we look at the part (b) of the theorem, the more complicated part, and highlight two major changes without repeating the whole proof of Proposition 2.

First, the convergence in (153) of the "non-delayed" version of the arrival process is modified as follows:

$$\begin{aligned}\hat{E}_r^{0,m_k}(t)&=\frac{1}{km_k}\big(E^{0,k}(k^2m_kt)-\lambda_r^kk^2m_kt\big)\\&=\frac{1}{\sqrt{m_k}}\cdot\frac{E^{0,k}(k^2m_kt)-\lambda_r^kk^2m_kt}{\sqrt{k^2m_k}}\to0\qquad\text{u.o.c. as }k\to\infty.\end{aligned}$$

The above convergence is u.o.c. because the limit is a deterministic continuous function (actually the limit is zero here). Then, we can show

$$\hat{E}_r^{m_k}(t_0^k+t)-\hat{E}_r^{m_k}(t_0^k)\to 0 \quad \text{u.o.c. as } k\to\infty.$$

Similarly, the convergence in (159) becomes

$$\hat{S}_r^{m_k}(\tilde{D}_r^{m_k}(t_0^k+t))-\hat{S}_r^{m_k}(\tilde{D}_r^{m_k}(t_0^k))\to 0 \quad \text{u.o.c. as } k\to\infty.$$

where $\tilde{D}_r^{m_k}(t)=D_r^k(k^2m_kt)/(k^2m_k)$ is another variation of fluid scaling, while the convergence in (158) remains intact. Hence, the convergence of the free process in (160) becomes

$$\hat{X}_r^{m_k}(t_0^k+t)-\hat{X}_r^{m_k}(t_0^k)\to \hat{x}_r(t):=\theta_r t. \tag{192}$$

In contrast to (160), the (driftless) Brownian motion component vanishes here, and thereafter in modifying the proofs of Theorems 1 and 2, the process $\hat{X}(t)$ should be understood as in (192); in other words, the processes $(\hat{W}(t),\hat{X}(t),\hat{Y}(t),\hat{Z}(t))$ referred to in the proofs of Theorem 1 and Proposition 2 are now replaced by $(\hat{w}(t),\hat{x}(t)=\theta t,\hat{y}(t),\hat{z}(t))$ defined in (39-44).

Second, the rescaling of the workload process in (161) becomes: for $u\in[0,T]$ and $j\le k\delta/T$,

$$\bar{W}^{k,j}(u)=\hat{W}^{m_k}(\tau+\frac{jT+u}{k})=\frac{1}{km_k}W^k((k^2m_k\tau+km_kjT)+km_ku).$$

The processes, $\bar{\Xi}^{k,j}(u)$, $\bar{U}^{k,j}(u)$, $\bar{V}^{k,j}(u)$, $\bar{N}^{k,j}(u)$ and $\bar{Y}^{k,j}(u)$, are similarly modified.

Appendix C: Proofs for §3 This main task here is to prove Lemma 6. While the main idea of the proof follows the proof of the diffusion limit theorems in Propositions 2 and 3, the version with a tight sequence of initial states, there are several key modifications, which we now highlight before presenting the proof.

First, the "free process" $\hat{X}^k(t)$ in the proof of the diffusion limit is replaced by $k(\rho^k-\rho)t$. Other hat-processes (e.g. $\hat{Y}^k(t)$) are changed to the corresponding lower-case processes (e.g., $\hat{y}^k(t)$). In addition, the allocation $\Lambda(\cdot)$ in the original networks is changed to $\bar{\Lambda}(\cdot)$ in the corresponding fluid networks (refer to their definitions in (3, 132)), which actually does not affect the key properties used in the proof. Specifically, Lemma 19 continues to hold if $\Lambda(n)$ is replaced by $\bar{\Lambda}(n)$. To see this, note the difference between $\Lambda(n)$ and $\bar{\Lambda}(n)$: $\Lambda_r(n)=0$ versus $\bar{\Lambda}_r(n)=\rho_r$, when $n_r=0$. Hence, it suffices to show the difference does not affect the conclusion stated in the lemma. Recall that ℓ^* is the link satisfying the condition in (142), and that $g_{\ell^*}^Tw>\epsilon>\sigma_1$. Consider any r such that $a_{\ell^*r}>0$. We have $r\in\mathcal{R}_{1,\sigma_1}$, since $\ell^*\in\mathcal{L}_{1,\sigma_1}$. Hence,

$$\begin{aligned}
w_r &= \sum_\ell b_ra_{\ell r}\pi_\ell+\delta_r=b_ra_{\ell^*r}\pi_{\ell^*}+\sum_{\ell\neq\ell^*}b_ra_{\ell r}\pi_\ell+\delta_r\\
&\ge b_ra_{\ell^*r}\sigma_1-\sigma\sum_{\ell\neq\ell^*}b_ra_{\ell r}-\kappa_1\sigma\ge b_{\min}a_{\min}\sigma_1-\left(\sum_{\ell,s}b_sa_{\ell s}+\kappa_1\right)\sigma\\
&> 0,
\end{aligned}$$

where the last inequality holds since σ can be made sufficiently small (relative to σ_1). Then, the above implies $\Lambda_r(n)=\bar{\Lambda}_r(n)$. Consequently, we have $A_{\ell^*}\bar{\Lambda}(n)=A_{\ell^*}\Lambda(n)=c_{\ell^*}$.

Second, we need to modify Lemma 17 to account for the residuals $\bar{u}^k(1)$ and $\bar{v}^k(1)$, which takes the following form (the detailed arguments omitted as they follow closely the proof of Lemma 17): For the sequence of $\hat{w}^k(t)$ (with uniformly bounded initial states particularly) considered in Lemma

6, there exist a time τ and a constant M, both independent of k, such that the following holds when k is sufficiently large:

$$\big(te, \bar{d}^k(t)\big) \geq \frac{1}{k}\left(\bar{u}^k(1), \bar{v}^k(1)\right), \text{ for } t \geq \frac{1}{k}\tau; \tag{193}$$
$$|\hat{w}^k(\tau/k)| \leq M. \tag{194}$$

That is, for the k-th network $\hat{w}^k(t)$, the residuals will have no effect after the time τ/k; and moreover, the workloads by that time are uniformly bounded by a constant that depends on network parameters only. Hence, it suffices to prove Lemma 6 for the case of setting the residuals $\bar{u}^k(1)$ and $\bar{v}^k(1)$ to zero, particularly for the equation in (56); and we shall do so below.

Third, the fluid limit and the uniform attraction property must be re-developed. This can be done via adapting the proofs of their stochastic counterparts in establishing the diffusion limit (i.e., Theorem 1 and Proposition 2). The details are presented in Lemmas 25-26 below.

Consider the time interval $[\tau, \tau+\delta]$, where $\tau \geq 0$ and $\delta > 0$. Let $T > 0$ be a fixed time length. Define for $u \geq 0$ and $j = 0, \cdots, k\delta/T$,

$$\bar{w}^{k,j}(u) = \hat{w}^k(\tau + \frac{jT+u}{k}).$$

Similarly, we define $\bar{n}^{k,j}(u)$, and $\bar{d}^{k,j}(u)$, etc., as the fluid-scaled "magnifiers" of the corresponding processes for the time interval $[\tau + jT/k, \tau + (j+1)T/k]$.

Lemma 25 Let j_k be an integer for each k. Suppose the L_1 norm of $\bar{w}^{k,j_k}(0)$ is bounded: $|\bar{w}^{k,j_k}(0)| = \sum_{r\in\mathcal{R}} \bar{w}_r^{k,j_k}(0) \leq M$ for all k and for some constant M. Then, for any subsequence of k, there exists a further subsequence, denoted $\mathcal{K}$, such that, along $\mathcal{K}$,

$$\big(\bar{w}^{k,j_k}(t), \bar{d}^{k,j_k}(t)\big) \to \big(\bar{w}(t), \bar{d}(t)\big) \qquad \text{u.o.c.},$$

where the ("fluid") limit is Lipschitz continuous and satisfies (129-132) (with $(\bar{u}(1), \bar{v}(1)) = 0$). Consequently, $\bar{w}(t)$ satisfies all the conclusions in Proposition 18 (with $\tau = 0$), assuming the heavy traffic condition in (11) holds.

The above is a special case of Proposition 4.2 in [57]: replacing the the renewal processes there by the (deterministic) λt and μt. (Also note that the allocation here is $\bar{\Lambda}(n)$ instead of $\Lambda(n)$, but this will not affect the proof as $\bar{\Lambda}_r(n) = \Lambda_r(n)$ for $n_r > 0$.)

Lemma 26 Consider the time interval $[\tau, \tau+\delta]$, with $\tau \geq 0$ and $\delta > 0$; let $C = |\theta|\delta$; and suppose

$$\sup_k |\hat{w}^k(\tau)| \leq M, \tag{195}$$

for some constant $M \geq 0$. Let $\epsilon > 0$ be any given (small) number. Then, there exists a sufficiently large T such that, for sufficiently large k, the following results hold for all *positive* integers $j < k\delta/T$ (i.e., excluding $j = 0$):

(a) (uniform attraction)

$$d^{fp}(\bar{w}^{k,j}(u)) \leq \epsilon, \qquad \text{for all } u \in [0, T]; \tag{196}$$

(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)

$$|\bar{w}^{k,j}(u)| \le \kappa_w |\hat{w}^k(\tau)| + \kappa_x(C+\epsilon) \le \kappa := \kappa_w M + \kappa_x(C+\epsilon), \quad \text{for all } u \in [0,T], \tag{197}$$

where κ_w is a positive constant specified in Proposition 18, and κ_x is also a positive constant that depends only on network parameters.

The above lemma is a straightforward adaptation of Lemma 23, replacing processes in the original stochastic networks (e.g., $\hat{W}^k(t)$, $\hat{Y}^k(t)$, $\bar{W}^{k,j_k}(u)$, $\bar{Y}^{k,j_k}(u)$, $\hat{X}^k(t)$, and $\Lambda(n)$) by their fluid counterparts (e.g., $\hat{w}^k(t)$, $\hat{y}^k(t)$, $\bar{w}^{k,j_k}(u)$, $\bar{y}^{k,j_k}(u)$, $\hat{x}^k(t) = \hat{w}^k(0) + k(\rho^k - \rho)t$, and $\bar{\Lambda}(n)$). Its proof also parallels closely that of Lemma 23, and hence omitted.

Proof (of Lemma 6). Keeping in mind the boundedness condition in the lemma ($|\hat{w}^k(0)| \le 1$), we know that the condition in (195) is satisfied with $\tau = 0$ and $M = 1$. Hence, we can apply Lemma 26, with $\tau = 0$, any $\delta > 0$ and $M = 1$, to the processes $\hat{w}^k(t)$. Pick any $\delta' \in (0,\delta)$. From the definition of t_0^k, we have for sufficiently large k and for any $t \in [0,\delta']$,

$$t_0^k + t = \frac{jT+u}{k}, \quad \text{for some } u \in [0,T] \text{ and } 1 \le j < k\delta/T.$$

Therefore, we can write

$$\hat{w}^k(t_0^k + t) = \bar{w}^{k,j}(u), \quad \text{for some } u \in [0,T] \text{ and } 1 \le j < k\delta/T.$$

Then, Lemma 26(a) translates to: for any $\epsilon > 0$ and sufficiently large k, the following holds,

$$d^{fp}(\hat{w}^k(t_0^k + t)) \le \epsilon, \quad \text{for } t \in [0,\delta'].$$

Therefore, we have

$$d^{fp}(\hat{w}^k(t_0^k + t)) \to 0 \quad \text{u.o.c of } t \ge 0, \text{ as } k \to \infty. \tag{198}$$

By the property (c) in Lemma 26, we have

$$|\hat{w}^k(t_0^k)| \le \kappa_w |\hat{w}^k(0)| + \kappa_x(C+\epsilon) \le \kappa_w + \kappa_x(C+\epsilon), \tag{199}$$

and hence

$$\limsup_{k\to\infty} |\hat{w}^k(t_0^k)| \le \kappa_w + \kappa_x(C+\epsilon).$$

Observe that the left hand side is independent of ϵ and C, and that $C(= |\theta|\delta)$ can be made arbitrarily small by choosing a sufficiently small δ from the beginning. Thus, the above inequality can be strengthened as

$$\limsup_{k\to\infty} |\hat{w}^k(t_0^k)| \le \kappa_w. \tag{200}$$

Consequently, for any subsequence of k, there exists a further subsequence $\mathcal{K}$ such that

$$\hat{w}^k(t_0^k) \to \hat{w}(0) \quad \text{as } k \to \infty \text{ along } \mathcal{K},$$

where due to (198) and (200), the state $\hat{w}(0)$ satisfies:

$$\hat{w}(0) \in \mathcal{W} \ (d^{fp}(\hat{w}(0)) = 0), \quad \text{and} \quad |\hat{w}(0)| \le \kappa_w.$$

Multiplying both sides of (56) by h_m^T yields

$$h_m^T \hat{w}^k(t) = h_m^T(\hat{w}^k(0) - \rho^k \bar{u}^k(1) + \bar{v}^k(1)) + h_m^T k(\rho^k - \rho)t + \hat{z}_m^k(t),$$

where we have applied (193) to simplify the terms involving the residuals. (One may choose to ignore these residual terms following our comments around (193), if one wish to.) The above equality implies the following,

$$h_m^T \hat{w}^k(t_0^k + t) = h_m^T \hat{w}^k(t_0^k) + h_m^T k(\rho^k - \rho)t + (\hat{z}_m^k(t_0^k + t) - \hat{z}_m^k(t_0^k)).$$

From the convergence in (198) and the definition of d^{fp} in (22), we have $h_m^T \hat{w}^k(t_0^k + t) \to 0$ as $k \to \infty$ (u.o.c. of $t \geq 0$). Therefore, letting $k \to \infty$ in the above equality yields,

$$\hat{z}_m^k(t_0^k + t) - \hat{z}_m^k(t_0^k) \to \hat{z}_m(t) := -h_m^T \theta t, \quad \text{u.o.c. of } t \geq 0.$$

The process $\hat{y}^k(t)$, for each k, is nondecreasing and continuous (refer to (58)). Moreover, the property (c) in Lemma 26 implies that $\hat{w}^k(t)$, and thus $\hat{y}^k(t)$, are uniformly bounded on any compact set of time t. (We can choose $\tau = 0$ and any δ in the lemma.) Hence, we are guaranteed that for any subsequence of $\mathcal{K}$ there exists a further subsequence, denoted by $\mathcal{K}'$, such that $(\hat{y}^k(t_0^k + t) - \hat{y}^k(t_0^k))$ converge along $\mathcal{K}'$ to a limit $\hat{y}(t)$, which is nondecreasing, RCLL, and is finite for all $t \geq 0$. Note that $\hat{y}(t)$ is continuous for almost all time t, and that as yet this convergence is guaranteed only for those times t at which $\hat{y}(t)$ is continuous. Consequently, we have, from (56),

$$\begin{aligned} \hat{w}^k(t_0^k + t) &= \hat{w}^k(t_0^k) + k(\rho^k - \rho)t + BG(\hat{y}^k(t_0^k + t) - \hat{y}^k(t_0^k)) + BH(\hat{z}^k(t_0^k + t) - \hat{z}^k(t_0^k)) \\ &\to \hat{w}(t) = \hat{w}(0) + \theta t + BG\hat{y}(t) + BH\hat{z}(t), \end{aligned}$$

along the same convergent subsequence $\mathcal{K}'$. Note that this convergence holds for those times t at which $\hat{y}(t)$ is continuous, and that $\hat{w}(t)$ is finite for all $t \geq 0$ as well.

From what has been established above, along with Lemma 26, it can be directly verified that the limits, $\hat{w}(t)$, $\hat{y}(t)$ and $\hat{z}(t)$, jointly satisfy (39-44). (In particular, the convergence in (198) implies $d^{fp}(\hat{w}(t)) = 0$, which gives (40) and (43).) Since $x(u) = \theta u$ is a continuous function, the oscillation inequality in Lemma 21 implies that $\hat{w}(t)$ and $\hat{y}(t)$ are also continuous.

Having proved that the convergence, along the subsequence $\mathcal{K}'$, to the limit $(\hat{w}(t), \hat{y}(t), \hat{z}(t))$ holds for all t, and that the limit is continuous and satisfies all the requirements in (39)-(44), we can invoke the uniqueness of the solution to the DCP in (39)-(44) (refer to Proposition 20) to conclude that the u.o.c. convergence holds for the subsequence $\mathcal{K}$. □

Appendix D: Proofs for §4

D.1. Proof of Lemma 9 As a preparation, we first turn the moment condition in (13) to the p-th moment condition on the non-delayed (i.e., renewal) part of the arrival and service processes (e.g., Theorem 4 (in the appendix 1) of [40]): for some $p \geq 2$, the following inequality holds uniformly for all k,

$$\mathsf{E} \sup_{0 \leq s \leq t} \sum_{r \in \mathcal{R}} \left(|E_r^{0,k}(s) - \lambda_r^k s|^p + |S_r^{0,k}(s) - \mu_r^k s|^p \right) \leq \kappa(1 + t^{\frac{p}{2}}), \quad t \geq 0, \tag{201}$$

which implies the following, more convenient formulation,

$$\mathsf{E} \sup_{0 \leq s \leq t} \sum_{r \in \mathcal{R}} \left(|\hat{E}_r^{0,k}(s)|^p + |\hat{S}_r^{0,k}(s)|^p \right) \leq \kappa(1 + t^{\frac{p}{2}}), \quad t \geq 0. \tag{202}$$

This version of p-th moment condition can be applied more directly below. It is also used in similar studies in [6, 34, 35].

We prove Lemma 9 now.

From the bounded workload condition (72), there exists a constant $\kappa_1 > 0$ such that the following condition holds,

$$\sup_{0\le s\le t} |\hat{W}^k(s)| \le \kappa_1 \left(|\hat{W}^k(0)| + \sup_{0\le s\le t} \sum_{r\in\mathcal{R}} \left(\nu_r^k |\hat{E}_r^k(t)| + \nu_r^k |\hat{S}_r^k(\tilde{D}_r^k(t))| + k|\rho_r^k - \rho_r| t \right) \right).$$

Recall that $\tilde{D}_r^k(t) = D_r^k(k^2 t)/k^2$. Then, we have, due to the capacity constraint,

$$\tilde{D}_r^k(t) = \frac{1}{k}\int_0^{k^2 t} \Lambda_r(N^k(s))ds \le \max_\ell c_\ell t,$$

which implies

$$\sup_{0\le s\le t} |\hat{S}_r^k(\tilde{D}_r^k(t))| \le \sup_{0\le s\le t} |\hat{S}_r^k(\max_\ell c_\ell t)|$$

From (10), we know that $\{\nu_r^k\}$ and $\{k|\rho_r^k - \rho_r|\}$ are uniformly bounded by some constant κ' over all k and r.

Also, we observe the effect of residuals in the arrival and service processes by examining their definitions in (4). If $t < u_r^k(1)/k^2$, we have $\hat{E}_r^k(t) = -\lambda_r^k k t$; otherwise, we have

$$\hat{E}_r^k(t) = \hat{E}_r^{0,k}(t - u_r^k(1)/k^2) + \frac{1}{k} - \lambda_r^k \frac{u_r^k(1)}{k}.$$

Combining these two cases yields

$$\sup_{0\le s\le t} |\hat{E}_r^k(s)| \le \sup_{0\le s\le t} |\hat{E}_r^{0,k}(s)| + \frac{1}{k} + \lambda_r^k \frac{u_r^k(1)}{k}.$$

Similarly, we have

$$\sup_{0\le s\le t} |\hat{S}_r^k(t)| \le \sup_{0\le s\le t} |\hat{S}_r^{0,k}(t)| + \frac{1}{k} + (\nu_r^k)^{-1} \frac{v_r^k(1)}{k}.$$

Hence, from the above estimates and simple algebra, we have for some constant κ_2

$$\sup_{0\le s\le t} |\hat{W}^k(s)| \le \kappa_2 \left(|\hat{\Xi}^k(0)| + t + 1 + \sum_{r\in\mathcal{R}} \left(\sup_{0\le s\le t} |\hat{E}_r^{0,k}(t)| + \sup_{0\le s\le t} |\hat{S}_r^{0,k}(\max_\ell c_\ell t)| \right) \right).$$

The above implies, for some constant κ_3,

$$\sup_{0\le s\le t} |\hat{W}^k(s)|^p \le \kappa_3 \left(|\hat{\Xi}^k(0)|^p + t^p + 1 + \sum_{r\in\mathcal{R}} \left(\sup_{0\le s\le t} |\hat{E}_r^{0,k}(t)|^p + \sup_{0\le s\le t} |\hat{S}_r^{0,k}(\max_\ell c_\ell t)|^p \right) \right). \quad (203)$$

Applying the p-th moment condition in (202), we have for some constants κ_4 and κ,

$$\mathsf{E} \sup_{0\le s\le t} |\hat{W}^k(s)|^p \le \kappa_3 \left(M^p + t^p + 1 + \kappa_4 (1 + t^{\frac{p}{2}}) \right) \le \kappa \left(M^p + 1 + t^p \right).$$

The proofs of (b) and (c) are similar, and we prove (c) below only. From (203), we have

$$\left|\frac{\hat{W}^k(m_k t)}{m_k}\right|^p \leq \kappa_3\left(1+t^p+\frac{1}{m_k^p}\sum_{r\in\mathcal{R}}\left(\sup_{0\leq s\leq m_k t}|\hat{E}_r^{0,k}(s)|^p+\sup_{0\leq s\leq m_k t}|\hat{S}_r^{0,k}(\max_{\ell} c_\ell s)|^p\right)\right). \quad (204)$$

For convenience, denote

$$\xi_r^k=\frac{1}{m_k^p}\sup_{0\leq s\leq m_k t}|\hat{E}_r^{0,k}(s)|^p \quad\text{and}\quad \eta_r^k=\frac{1}{m_k^p}\sup_{0\leq s\leq m_k t}|\hat{S}_r^{0,k}(\max_{\ell} c_\ell s)|^p$$

Applying the p-th moment condition in (202) yields, for some constant κ_5, as $k\to\infty$,

$$\mathsf{E}\xi_r^k \leq \frac{\kappa_5(1+(m_k t)^{p/2})}{m_k^p}\to 0,$$

and similarly $\mathsf{E}\eta_r^k\to 0$. Now, for any (small) $\epsilon>0$, we can find a sufficiently large K such that $\mathsf{E}\xi_r^k<\epsilon$ for all $k>K$, and then we have

$$\begin{aligned}\sup_k \mathsf{E}\xi_r^k 1_{\{\xi_r^k>a\}} &\leq \max\left\{\sup_{k\leq K}\mathsf{E}\xi_r^k 1_{\{\xi_r^k>a\}},\ \sup_{k>K}\mathsf{E}\xi_r^k\right\}\\ &\leq \max\left\{\sup_{k\leq K}\mathsf{E}\xi_r^k 1_{\{\xi_r^k>a\}},\ \epsilon\right\}\to\epsilon, \quad \text{as } a\to\infty.\end{aligned}$$

That is, $\{\xi_r^k\}$ is uniformly integrable, and so is $\{\eta_r^k\}$. This integrability, along with the inequality in (204), implies the conclusion in (c). □

D.2. Proof of Lemma 10(b) Since $|\hat{\Xi}^k(0)|\leq m_k$, we can apply Proposition 3(b), with the sequence $\{t_0^k\}$ also given in the proposition. That is, for any subsequence of k, there exists a further subsequence, denoted by $\mathcal{K}$, such that the following weak convergence holds when $k\to\infty$ along $\mathcal{K}$:

$$\frac{1}{m_k}\hat{W}^k(m_k(t_0^k+t))\Rightarrow\hat{w}(t)$$

where the limit $\hat{w}(t)$ (along with suitable $\hat{y}(t)$ and $\hat{z}(t)$) follows the specifications in (39-44). Furthermore, according to (46) with M replaced by κ_w, the limit satisfies $|\hat{w}(0)|\leq\kappa_w$ with probability one. (Recall, κ_w is a constant that depends only on network parameters; refer to Proposition 18(a).)

Under the stability of $\hat{w}(t)$ (which holds under the usual traffic condition in (66) for the case of resource-sharing network, as explained in Theorem 8(a)), the above limit satisfies

$$\hat{w}(t_0'+t)=0, \quad t\geq 0,$$

for some constant time $t_0'>0$. Therefore, we have from the above two displays: as $k\to\infty$ along $\mathcal{K}$,

$$\frac{1}{m_k}\hat{W}^k(m_k(t_0^k+t_0'+t))\Rightarrow\hat{w}(t_0'+t)\equiv 0.$$

Since the limit is unique and continuous (zero function) and $\mathcal{K}$ is a subsequence of an arbitrarily chosen subsequence of k, the above weak convergence is turned into a (u.o.c.) convergence along the full sequence of k with probability one: as $k\to\infty$ (along the full sequence),

$$\frac{1}{m_k}\hat{W}^k(m_k(t_0^k+t_0'+t))\to 0, \quad \text{u.o.c. of } t\geq 0.$$

Letting $t_0=t_0'+1$, the above implies the conclusion in Lemma 10(b). □

Acknowledgments. Heng-Qing Ye was supported in part by HK/RGC Grant 15508114. David Yao was supported in part by NSF Grants CMMI-0969328 and CMMI-1462495.

References

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

[2] Berman, A. and R.J. Plemmons, *Nonnegative Matrices in the Mathematical Science.* Academic Press, New York, 1979.

[3] Bonald, T. and L. Massoulie, Impact of Fairness on Internet Performance. *Proceedings of ACM Sigmetrics, 2001.*

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

[5] Bramson, M., Stability of Two Families of Queueing Networks and a Discussion of Fluid Limits. *Queueing Systems: Theory and Applications*, **23** (1998), 7-31.

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

[7] Chen, H. and J.G. Shanthikumar, Fluid Limits and Diffusion Approximations for Networks of Multi-Server Queues in Heavy Traffic. *Journal of Discrete Event Dynamic Systems*, **4** (1994), 269-291.

[8] Chen, H. and D.D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization.* Springer-Verlag, New York, 2001.

[9] Chen, H. and H.Q. Ye, Existence Condition for the Diffusion Approximations of Priority Multi-class Queueing Networks. *Queueing Systems: Theory and Applications,* **38** (2001), 435-470.

[10] Chen, H. and H. Zhang, Diffusion Approximations for Kumar-Seidman Network under a Priority Service Discipline. *Operations Research Letters*, **23** (1998), 171-181.

[11] Chen, H. and H. Zhang, A Sufficient Condition and a Necessary Condition for the Diffusion Approximations of Multiclass Queueing Networks under Priority Service Disciplines. *Queueing Systems, Theory and Applications,* **34** (2000b), 237-268.

[12] 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.

[13] Dai, J.G., A.B. Dieker, and X. Gao, Validity of Heavy-traffic Steady-state Approximations in Many-server Queues with Abandonment. *Queueing Systems: Theory and Applications*, to appear.

[14] Dai, J.G. and W. Lin, Asymptotic Optimality of Maximum Pressure Policies in Stochastic Processing Networks. *Annals of Applied Probability*, **18** (2008), 2239-2299.

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

[16] Dai, J.G. and J.H. Vande Vate, The stability of two-station multi-type fluid networks. *Operations Research*, **48** (2000), 721-744.

[17] 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.

[18] de Veciana, G., T.J. Lee, and T. Konstantopoulos, Stability and Performance Analysis of Networks Supporting Elastic Services. *IEEE/ACM Transactions on Networking*, **9** (2001), 2-14.

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

[20] Eryilmaz, A. and R. Srikant, Asymptotically Tight Steady-State Queue-length Bounds Implied by Drift Conditions. *Queueing Systems: Theory and Applications*, **72** (2012), No.3-4, 311359.

[21] Gamarnik D. and D. Goldberg, On the Rate of Convergence to Stationarity of the M/M/N Queue in the Halfin-Whitt Regime. *Annals of Applied Probability*, **23** (2013), 1879-1912.

[22] Gamarnik, D. and A. Stolyar, Multiclass Multiserver Queueing System in the Halfin-Whitt Heavy Traffic regime: Asymptotics of the Stationary Distribution. *Queueing Systems: Theory and Applications*, **71** (2012), No.1-2, 25-51.

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

[24] Gromoll, H.C. and R.J. Williams, Fluid Model for a Data Network with α-Fair Bandwidth Sharing and General Document Size Distributions: Two Examples of Stability. *Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz* (Stewart N. Ethier, Jin Feng and Richard H. Stockbridge, eds.), Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2008, 253-265.

[25] Gromoll, H.C. and R.J. Williams, Fluid Limits for Networks with Bandwidth Sharing and General Document Size Distributions. *Annals of Applied Probability*, **19** (2009), 243-280.

[26] 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), No. 1, 121-162.

[27] Harrison, J.M., The Heavy Traffic Approximation for Single Server Queues in Series. *Journal of Applied Probability*, **10** (1973), 613-629.

[28] Harrison, J.M., The Diffusion Approximation for Tandem Queues in Heavy Traffic. *Advances in Applied Probability,* **10** (1978), 886-905.

[29] Harrison, J.M., Brownian Models of Open Processing Networks: Canonical Representation of Workload. *Annals of Applied Probability*, **10** (2000), No. 1, 75-103. Correction: **13** (2003), No. 1, 390-393.

[30] Harrison, J.M., A Broader View of Brownian Networks. *Annals of Applied Probability,* **13** (2003), No. 3, 1119-1150.

[31] Harrison, J.M., C. Mandayam, D. Shah, and Y. Yang, Approaching HGI performance in resource sharing networks. Working paper (2013).

[32] Kang, W.N., Kelly, F.P., Lee, N.H. and Williams, R.J., State Space Collapse and Diffusion Approximation for a Network Operating under a Fair Bandwidth Sharing Policy. *Annals of Applied Probability*, **19** (2009), No. 5, 1719-1780.

[33] Kaspi, H. and Mandelbaum, A., Regenerative Closed Queueing Networks. *Stochastics and Stochastic Reports*, **39**, 230-258, 1992.

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

[35] Katsuda, T., Heavy-Traffic Approximation for Stationary Distribution in a Multi-class Single-Server Queue. Submitted for publication, 2011.

[36] Kelly, F.P., Loss Networks. *Annals of Applied Probability,* **1** (1991), 319-378.

[37] Kelly, F.P., A. Maulloo and D. Tan, Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability. *Journal of the Operational Research Society,* **49** (1998), 237-252.

[38] Kelly, F.P. and R.J. Williams, Fluid Model for a Network Operating under a Fair Bandwidth-Sharing Policy. *Annals of Applied Probability,* **14** (2004), 1055-1083.

[39] Kelly, F.P. and R.J. Williams, Heavy Traffic on a Controlled Motorway. *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman* (Editors N.H. Bingham and C.M. Goldie), Cambridge University Press, 2010.

[40] Krichagina, E.V. and M.I. Taksar, Diffusion approximation for GI/G/1 controlled queues. *Queueing Systems: Theory and Applications*, **12** (1992), 333-368.

[41] Kumar, P.R. and T.I. Seidman, Dynamic Instabilities and Stabilization Methods in Distributed Real-time Scheduling of Manufacturing Systems. *IEEE Transactions on Automatic Control*, **35** (1990), 289-298.

[42] Mandelbaum, A. and A.L. Stolyar, Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized $c\mu$-Rule. *Operations Research*, **52** (2004), 836-855.

[43] Massoulie, L. and J.W. Roberts, Bandwidth Sharing and Admission Control for Elastic Traffic. *Telecommunication Systems,* **15** (2000), 185-201.

[44] Meyn, S. and R.L. Tweedie, *Markov Chain and Stochastic Stability.* Springer-Verlag, London, 1993.

[45] Mo, J. and J.C. Walrand, Fair End-to-End Window-based Congestion Control. *IEEE/ACM Transaction on networking*, **8** (2000), 556-567.

[46] Ramanan, K. and M. Reiman, Fluid and Heavy Traffic Limits of a Generalized Processor Sharing Model. *Annals of Applied Probability*, **13** (2003), 100-139.

[47] Rybko, A.N. and A. L. Stolyar, Ergodicity of Stochastic Processed Describing the Operations of Open Queueing Networks. *Problemy Peredachi Informatsii*, **28** (1992), 2-26.

[48] Shah, D., J.N. Tsitsiklis and Y. Zhong, Qualitative Properties of α-Fair Policies in Bandwidth-Sharing Networks. *Annals of Applied Probability*, **24** (2014), 76-113.

[49] Stolyar, A.L., On the Stability of Multi-class Queueing Network: a Relaxed Sufficient Condition via Limiting Fluid Processes. *Markov Processes and Related Fields*, **1** (1995), No. 4, 491-512.

[50] Stolyar, A.L., Max-Weight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic. *Annals of Applied Probability,* **14** (2004), 1-53.

[51] W. Szczotka and F. P. Kelly, Asymptotic Stationarity of Queues in Series and the Heavy Traffic Approximation. *Annals of Probability*, **18** (1990), No. 3, 1232-1248.

[52] Tezcan, T., Optimal Control of Distributed Parallel Server Systems under the Halfin and Whitt Regime. *Mathematics of Operations Research*, **33** (2008), 51-90.

[53] Williams, R.J., An Invariance Principle for Semimartingale Reflecting Brownian Motions in an Orthant. *Queueing Systems: Theory and Applications*, **30** (1998), 5-25.

[54] Williams, R.J., Diffusion Approximations for Open Multi-class Queueing Networks: Sufficient Conditions Involving State Space Collapse. *Queueing Systems: Theory and Applications*, **30** (1998), No. 1-2, 27-88.

[55] Ye, H.Q., Stability of Data Networks Under An Optimization-Based Bandwidth Allocation. *IEEE Transactions on Automatic Control*, **48** (2003), No. 7, 1238-1242.

[56] Ye, H.Q. and H. Chen, Lyapunov Method for the Stability of Fluid Networks. *Operations Research Letters*, **28** (2001), 125-136.

[57] Ye, H., J. Ou and X. Yuan, Stability of Data Networks: Stationary and Bursty Models. *Operations Research*, **53** (2005), No. 1, 107-125.

[58] Ye, H. and D.D. Yao, Heavy Traffic Optimality of a Stochastic Network under Utility-Maximizing Resource Control. *Operations Research*, **56** (2008), No. 2, 453-470.

[59] Ye, H.Q. and D.D. Yao, Utility-Maximizing Resource Control: Diffusion Limit and Asymptotic Optimality for a Two-Bottleneck Model. *Operations Research*, **58** (2010), 613-623.

[60] 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.