Adv. Appl. Prob. **45,** 645–672 (2013)
Printed in Northern Ireland
© *Applied Probability Trust* 2013

HEAVY-TRAFFIC LIMITS FOR A MANY-SERVER QUEUEING NETWORK WITH SWITCHOVER

GUODONG PANG,* *Pennsylvania State University*
DAVID D. YAO,** *Columbia University*

Abstract

We study a multiclass Markovian queueing network with switchover across a set of many-server stations. New arrivals to each station follow a nonstationary Poisson process. Each job waiting in queue may, after some exponentially distributed patience time, switch over to another station or leave the network following a probabilistic and state-dependent mechanism. We analyze the performance of such networks under the many-server heavy-traffic limiting regimes, including the critically loaded quality-and-efficiency-driven (QED) regime, and the overloaded efficiency-driven (ED) regime. We also study the limits corresponding to mixing the underloaded quality-driven (QD) regime with the QED and ED regimes. We establish fluid and diffusion limits of the queue-length processes in all regimes. The fluid limits are characterized by ordinary differential equations. The diffusion limits are characterized by stochastic differential equations, with a piecewise-linear drift term and a constant (QED) or time-varying (ED) covariance matrix. We investigate the load balancing effect of switchover in the mixed regimes, demonstrating the migration of workload from overloaded stations to underloaded stations and quantifying the load balancing impact of switchover probabilities.

Keywords: Many-server queue; multiclass; Markovian; time-varying arrival; state-dependent switchover; customer abandonment; heavy traffic; QED regime; ED regime; mixed regime; fluid limit; diffusion limit; multidimensional (piecewise-linear) Ornstein–Uhlenbeck process

2010 Mathematics Subject Classification: Primary 60K25
Secondary 60F17; 90B22

1. Introduction

We study a multiclass Markovian queueing network with a set of nodes, each representing a service facility with multiple parallel servers. Each node serves a class of jobs, which follows a nonstationary Poisson arrival process, and requires independent and identically distributed exponential service times. The service discipline is first-come–first-served (FCFS), and each node keeps its own queue. A job waiting in queue may, after some exponentially distributed 'impatience' time, switch over to another node or leave the system ('abandonment') following a probabilistic and state-dependent mechanism. Figure 1 depicts such a network with two nodes.

Waiting customers switching queues is a commonplace phenomenon in many service systems. This, however, is not always due to customer impatience: in some application contexts, switching queues could be a necessity or even part of the design. For instance, in many hospitals,

Received 1 March 2011; revision received 23 November 2012.
* Postal address: The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802, USA. Email address: gup3@psu.edu
** Postal address: Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA. Email address: yao@columbia.edu

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

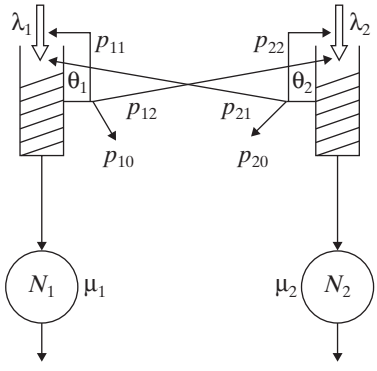

FIGURE 1: Multiclass many-server queueing network with switchovers.

there is a standard practice to switch an emergency patient, who has been waiting for an available bed from a designated ward, to an alternative ward if the delay has reached a certain threshold. In this case, switchover is necessary. Switchover is also a means to deal with nonstationary demand patterns, since in most hospitals it is impossible to dynamically re-allocate beds across departments/wards. (This is in stark contrast against the dynamic staffing mechanism at call centers.) In wireless communications, the dynamic spectrum access (DSA) technology allows a client, who initially picks a server (frequency band) upon arrival, to sample the system state at random time points and migrate to another server—join the shortest queue, for instance. In this case, the switchover is essentially a load balancing protocol, designed to offset the uneven load (again, due to nonstationary arrivals), so as to achieve better resource utilization as well as improved service quality.

We study this class of networks in the many-server heavy-traffic limiting regimes, the *critically loaded* quality-and-efficiency-driven (QED) regime (i.e. the traffic intensity equals the service capacity), the *overloaded* efficiency-driven (ED) regime, and the mixing of the *underloaded* quality-driven (QD) regime with the QED and ED regimes. In all limiting regimes, the arrival rate and the number of servers are scaled up to ∞ in a suitable manner. Thus, our model is relevant to applications where there are a large number of servers and a high rate of demand arrivals (when measured against service requirements). The healthcare and telecommunications examples above certainly fall within this range of applications.

For both QED and ED regimes, we establish the fluid and diffusion limits of the queue-length processes. The fluid limits (Theorems 1 and 4) are characterized by ordinary differential equations (ODEs). The diffusion limits are characterized by stochastic differential equations (SDEs), with a piecewise-linear drift term and a constant diagonal covariance matrix in the QED regime (Theorem 2). Under the ED regime, the SDE has a linear drift term, which has to be coupled with the fluid limit when the switchover is state dependent, and a time-varying covariance matrix (Theorem 5). For both regimes, we provide a characterization, via ODEs, for the mean and the covariance of the diffusion limits (Corollaries 2 and 5). Moreover, in the QED regime, we also establish the diffusion limit for the virtual waiting time process (Theorem 3).

As mentioned earlier, in many applications, there will be an unavoidable mismatch between supply and demand—service capacity and traffic intensity. Hence, it will be more useful and relevant to study mixed regimes. Here, we investigate all possible mixtures of the three regimes, QD, QED, and ED, focusing on the case when the arrival rates and the switchover probabilities are constant, and establish both fluid and diffusion limits (Theorems 6 and 7) for the

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

queue-length processes at all stations. Interestingly, the various mixtures exhibit qualitatively different behavior. In the mixed QD/QED regime, the steady state of the fluid limit is not affected by the switchover. However, the switchover from the QED stations to ED stations does affect the diffusion limit process, in particular, adding an extra drift term. In the mixed QED/ED regime, the switchover affects both the fluid and diffusion limits; in fact, it makes all stations overloaded, and, thus, the limits are the same as those in the ED regime. In the mixed QD/ED regime, there can be several possibilities: the QD stations can become overloaded or critically loaded, or remain underloaded, depending on the switchover probabilities. We characterize the fluid limit and its steady state and the diffusion limit that include all possibilities. Finally, in the mixed QD/QED/ED regime, the limits have the same characteristics as in the mixed QD/ED regime, i.e. adding QED to the latter does not qualitatively affect the regime.

A brief review of related papers in the literature is in order. Mandelbaum *et al.* (1998) studied many-server Markovian networks under the QED regime, with a general network configuration, time-varying arrival and service rates, and time-varying routeing probabilities. Our model, when specialized to the QED regime with constant switchover probabilities, can be regarded as an instance of theirs; in particular, the Jackson network model with abandonment in their Section 6. But we do allow the switchover probabilities to be state dependent, and also consider the ED regime, which they did not. More importantly, we consider the mixed QD, QED, and ED regimes, showing the impact of switchover. In addition, the methodologies are different: theirs is based on strong approximations, whereas we apply the martingale approach along with the continuous mapping theorem (CMT), *a la* Pang *et al*. (2007).

In an earlier work, Fleming *et al*. (1995) formulated a stochastic model of a mobile phone system, with one server pool supporting two types of calls with different impatience rates: beyond the patience limits, type-1 calls will quit the service and type-2 calls will choose to switch to type-1 or quit. The authors conjectured a heavy-traffic limit for the total number of calls in the system and the breakdown of waiting calls of either type. Our model is a generalization of theirs and our limiting results confirm their conjecture.

In some applications of the queueing model with switchovers, notably, wireless networks, two key issues are stability and load balancing. Examples include studies of how various switchover policies, both load oblivious (e.g. random local search) and load aware (e.g. join the shortest queue and its variants), lead to stability and how effective they can achieve load balancing We refer the reader to Ganesh *et al*. (2010), Bonald *et al*. (2009), Bramson *et al*. (2010), Simatos and Tibi (2010), and the references therein. In this regard, our study is quite different. First, by allowing switchover to include abandonment, stability is guaranteed in our model. On the other hand, we have a very general switchover model—the state-dependent switchover probabilities can handle both load-aware and load-oblivious policies as special cases. Second, our focus is on performance analysis in general, as opposed to load balancing in particular—to derive the fluid and diffusion limits so as to capture the impact of time-varying arrivals and state-dependent switchovers on system performance, such as congestion and delay.

In the queueing literature, the QED regime can be traced back to Erlang, who first described it back in 1924. Its formal analysis was done by Halfin and Whitt (1981) for a stand-alone many-server queue with general arrivals and exponential service times. Garnett *et al*. (2002) extended the analysis to include abandonment. This regime has been widely used to analyze the performance of call centers, e.g. Borst *et al*. (2004), and to support routeing and scheduling decisions in networks of many-server queues, e.g. Gurvich and Whitt (2009). The ED regime for the many-server queueing model but with the additional feature of abandonment was studied in Whitt (2004), and applied to call centers with unexpected overloads by Perry and Whitt (2013).

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

In the case of a single node, our network specializes to the queueing model with abandonment by reducing the switchover probabilities to abandonment probabilities. Studies of multiple service stations often assume that they all operate in the QED or ED regime. Here, we investigate the system behavior and the impact of switchover in the network when the stations are assumed to be in the mixed regimes.

The Markovian setting is crucial to our study, in exploiting the martingale approach to establish our limiting results. Specifically, our approach relies heavily on the CMT—that the multidimensional integral representations of the queueing processes are continuous in the Skorokhod J_1 topology (see Lemma 2, and the proofs of Theorems 2, 5, and 7). A non-Markovian setting for many-server heavy-traffic regimes will require a very different approach based on measure-valued processes; see, e.g. Kang and Ramanan (2010) and Kang and Pang (2013). This, however, does not diminish the value of our results. Indeed, with features such as nonstationary arrivals and state-dependent switchovers, we have greatly enriched the Markovian setup, bringing it closer to the reality of many applications highlighted earlier. Furthermore, because of these general features, the fluid and diffusion limits are far from trivial (such as convergence to 0 in some steady state or some invariant distribution as in more conventional heavy-traffic regimes)—they are derived below with considerable effort and expressed in the form of ODEs and SDEs.

The rest of the paper is organized as follows. We start with the model description in Section 2. The QED and the ED regimes are studied in Sections 3 and 4, respectively, and the mixed regimes in Section 5. Each section starts with a specification of the heavy-traffic condition that defines the regime, followed by the presentation of both fluid and diffusion limits, and related discussions and remarks on the results. Brief concluding remarks are summarized in Section 6, and all proofs are collected in Appendix A.

2. Model description

Here is a summary of the notation and regulations used in this paper. For $x, y \in \mathbb{R}$, $x^+ := \max\{x, 0\}$, $x^- := -\min\{x, 0\}$, $x \wedge y := \min\{x, y\}$, and $x \vee y := \max\{x, y\}$. Let '$\overset{\mathrm{D}}{\to}$' denote convergence in distribution. For a positive integer k, $D^k := D([0, \infty), \mathbb{R}^k)$ is the space of functions mapping $[0, \infty)$ into $\mathbb{R}^k$ that are right continuous in $[0, \infty)$ and have left limits in $(0, \infty)$. Write $D := D^1$. Let C^k be the subspace of continuous functions in D^k. The space D^k is endowed with the Skorokhod J_1 topology; see Billingsley (1999) and Whitt (2002) for background. Note that, when $x_n \to x$ in (D^k, J_1) as $n \to \infty$ and $x \in C^k$, the convergence is equivalent to the convergence in the topology of uniform convergence on bounded intervals. Define the vector norm $\|z\| := \sum_{i=1}^k |z_i|$ for any vector $z = (z_1, \ldots, z_k)^\top \in \mathbb{R}^k$ and the matrix norm $\|Z\| := \sum_{i,j=1}^k |Z_{ij}|$ for any matrix $Z \in \mathbb{R}^{k\times k}$. Since all norms are equivalent in the Euclidean vector space, we will use this particular norm in the proofs. We also define the norm $\|z\|_T := \sup_{0\le t\le T} \|z(t)\|$ for $z \in D^k$ and $0 < T < \infty$. For $z \in D$, the same notation $\|z\|_T$ is used without causing confusion. For $A, B \in \mathbb{R}^{k\times k}$ and $c \in \mathbb{R}^k$, AB and Ac denote matrix multiplication and matrix-vector multiplication.

We now describe our network model. There are K nodes (or service stations) in the network, indexed by $i = 1, \ldots, K$. Each node i represents a server pool, with N_i parallel servers, providing service to its own class of jobs, on an FCFS basis, and maintaining its own queue. The arrival process of class-i jobs is a nonstationary Poisson process with an intensity function $\lambda_i(t)$. Each job requires an independent, exponentially distributed service time of rate μ_i. Each job waiting in queue, if it does, has an independent, exponentially distributed patience time of rate θ_i, before switchover happens (to be specified immediately below).

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

Let $X_i := \{X_i(t) : t \geq 0\}$ denote the queue-length process at node i, $i = 1, \ldots, K$; specifically, $X_i(t)$ is the total number of class-i jobs, both in service and in queue, at time t. Denote by $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$ a vector of K-dimensional processes with sample paths in the space D^K. Let $\boldsymbol{P} := [p_{ij}(\cdot)]_{i,j=1}^K : \mathbb{R}_+^K \to [0,1]^{K\times K}$ be a matrix-valued function, where each component is a function of an $\mathbb{R}_+^K$-valued vector and takes values in $[0,1]$. Denote the space of such matrix functions as $\mathcal{M}^K$. We assume that, for each i, $\sum_{j=1}^K p_{ij}(\cdot) \leq 1$, $p_{ii} \in [0,1)$, and define $p_{i,0}(\cdot) := 1 - \sum_{j=1}^K p_{ij}(\cdot)$. If the patience time of a class-i job is reached at t, it will switch over to queue j, or leave the system, with a state-dependent probability $p_{ij}(\boldsymbol{X}(t-))$, or $p_{i0}(\boldsymbol{X}(t-))$, respectively. Assume that $\boldsymbol{X}(0-) = \boldsymbol{X}(0)$.

By counting the input and output quantities at each node, we have the following equivalent-in-distribution representation of the process $\boldsymbol{X}$ in terms of unit-rate Poisson processes:

$$\begin{aligned} X_i(t) = X_i(0) &+ A_i\left(\int_0^t \lambda_i(s)\,\mathrm{d}s\right) - S_i\left(\mu_i \int_0^t (X_i(s) \wedge N_i)\,\mathrm{d}s\right) \\ &+ \sum_{k=1,\,k\neq i}^K L_{k,i}\left(\theta_k \int_0^t p_{ki}(\boldsymbol{X}(s-))(X_k(s) - N_k)^+\,\mathrm{d}s\right) \\ &- \sum_{j=1,\,j\neq i}^K L_{i,j}\left(\theta_i \int_0^t p_{ij}(\boldsymbol{X}(s-))(X_i(s) - N_i)^+\,\mathrm{d}s\right) \\ &- L_{i,0}\left(\theta_i \int_0^t p_{i0}(\boldsymbol{X}(s-))(X_i(s) - N_i)^+\,\mathrm{d}s\right) \end{aligned} \tag{1}$$

for each $i = 1, \ldots, K$ and $t \geq 0$. Here A_i, S_i, and $L_{i,j}$ ($i = 1, \ldots, K$, $j = 0, 1, \ldots, K$) are mutually independent unit-rate Poisson processes, respectively representing the arrival, service-completion, and switchover (counting) processes. Specifically, the $L_{k,i}$ term counts the number of jobs switching into node i from node k, the $L_{i,j}$ term counts the number of jobs switching from node i to node j, and the $L_{i,0}$ term counts the number of jobs leaving the system from queue i after the impatience time.

We also let $\boldsymbol{V}(t) = (V_1(t), \ldots, V_K(t))^\top$ be the vector of virtual waiting time processes, where $V_i(t)$ is the potential waiting time of a hypothetical job arriving at the queue i at time t to be served at server pool i without switching over to other queues.

In the next two sections we will analyze the performance of such networks with switchovers in the two heavy-traffic limiting regimes, QED and ED regimes. In both regimes, we consider a sequence of such networks indexed by a scaling parameter $\lambda \in \mathbb{R}_+$, and let $\lambda \to \infty$. We write $\mathcal{X}^\lambda$ for any process $\mathcal{X} \in \{\boldsymbol{X}, \boldsymbol{V}\}$, $\boldsymbol{P}^\lambda$ for the switchover transition probability matrix, and N_i^λ for the number of servers at node i, while fixing the unit-rate Poisson processes A_i, $L_{i,j}$, and S_i, service rates μ_i, and switchover rates θ_i. Note that the arrival rate functions $\lambda_i(t)$, $i = 1, \ldots, k$, will also change along with λ, although we will omit the superscript λ in the notation.

3. The QED regime

In the QED regime, the system is critically loaded in heavy traffic; namely, the rate of work (service requirements) injected into each node is matched, in terms of the expectation, by the total service capacity at the node. The situation is more delicate with time-varying arrival rates. Additional conditions are needed such that the arrival rates do not fluctuate too much.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

Heavy-Traffic QED Assumptions. (i) The scaled number of servers satisfies the following condition: there exist positive constants ν_i and γ_i, $i = 1, \ldots, K$, such that

$$\frac{N_i^\lambda}{\lambda} \to \nu_i \quad \text{as } \lambda \to \infty \tag{2}$$

and

$$\frac{N_i^\lambda - \nu_i \lambda}{\sqrt{\lambda}} \to \gamma_i \in \mathbb{R} \quad \text{as } \lambda \to \infty. \tag{3}$$

(ii) The scaled arrival rate functions satisfy the following condition: for all $i = 1, \ldots, K$ and $T > 0$,

$$\sup_{0 \le t \le T} \left| \frac{1}{\lambda} \int_0^t \lambda_i(s)\, \mathrm{d}s - a_i t \right| \to 0 \quad \text{as } \lambda \to \infty, \tag{4}$$

where

$$a_i = \nu_i \mu_i \tag{5}$$

and

$$\sup_{0 \le t \le T} \left| \sqrt{\lambda} \left(\frac{1}{\lambda} \int_0^t \lambda_i(s)\, \mathrm{d}s - a_i t \right) - \int_0^t \xi_i(s)\, \mathrm{d}s \right| \to 0 \quad \text{as } \lambda \to \infty, \tag{6}$$

where $\xi_i : \mathbb{R}_+ \to \mathbb{R}$ is a deterministic function.

(iii) There exist $\lambda_i^+ > 0$ and $\lambda_i^- > 0$, $i = 1, \ldots, K$, possibly depending on λ, such that, when λ is large enough,

$$|\lambda_i^+ - \mu_i N_i^\lambda| \vee |\lambda_i^- - \mu_i N_i^\lambda| \le \varepsilon \tag{7}$$

holds for any small $\varepsilon > 0$. Moreover,

$$\sup_{0 \le t \le T} |\lambda_i(t+h) - \lambda_i(t)| \le \lambda_i^+ h, \qquad \inf_{0 \le t \le T} |\lambda_i(t+h) - \lambda_i(t)| \ge \lambda_i^- h \tag{8}$$

hold for any small $h > 0$.

Note that assumption (iii) implies that

$$\sup_{0 \le t \le T} \left| \int_0^t \lambda_i(s)\, \mathrm{d}s - \mu_i N_i^\lambda t \right| \to 0 \quad \text{as } \lambda \to \infty,\ i = 1, \ldots, K,$$

that is, the arrival rate matches the service capacity, as the scaling parameter $\lambda \to \infty$. In this sense, the system is critically loaded, i.e. the QED regime.

We should also expect that $\lambda^{-1}|\lambda_i^+ - \lambda_i^-| \to 0$ as $\lambda \to \infty$. Thus, in (4), a_i being constant is necessary and the rates match in the limit, $a_i = \nu_i \mu_i$ for each $i = 1, \ldots, K$.

As an example, consider a sinusoidal arrival rate function, $\lambda_i(t) = \bar{\lambda}_i + \delta_i \sin(\omega t)$, where $\bar{\lambda}_i$ is the average arrival rate, δ_i is the amplitude, and ω is the frequency. Our assumption requires that $\bar{\lambda}_i/\lambda \to a_i$ and $\delta_i/\lambda \to 0$ as $\lambda \to \infty$. Since $\lambda_i(t+h) - \lambda_i(t) = \delta_i \omega h + o(h^2)$, a suitable choice of ω will satisfy the above assumption.

In the special case of constant arrival rates, assumption (iii) is not needed; and assumption (ii) reduces to

$$\frac{\lambda_i}{\lambda} \to a_i = \nu_i \mu_i, \qquad \frac{\lambda_i - a_i \lambda}{\sqrt{\lambda}} \to \xi_i, \qquad i = 1, \ldots, K, \tag{9}$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

with the ξ_i being constants. These are standard heavy-traffic conditions for studying many-server queueing models in the QED regime; see, e.g. Gurvich and Whitt (2009). Specifically, $a_i = \nu_i \mu_i$ ensures the critical load condition at every node i, since by (9), $\lambda_i \approx a_i \lambda$ and $N_i^\lambda \approx \lambda \nu_i$ so that $\lambda_i \approx N_i^\lambda \mu_i$ for large λ. Consequently, under the fluid scaling, there is no queue in steady state, each class of jobs will be served by its own server pool; and hence, switchovers have no impact on the steady state of system performance. The other condition in (9) is required to establish the diffusion limits, along with the condition in (3). Note the latter is equivalent to $\sqrt{\lambda}(1 - \tilde{\rho}_i^\lambda) \to \tilde{\gamma}_i$ as $\lambda \to \infty$, where $\tilde{\rho}_i^\lambda := \lambda_i / N_i^\lambda \mu_i$ and $\tilde{\gamma}_i = \nu_i^{-1}(\gamma_i - \xi_i/\mu_i)$. Hence, this is similar to the usual heavy-traffic QED assumption in the single class case.

Define the fluid-scaled processes $\bar{\boldsymbol{X}}^\lambda := \boldsymbol{X}^\lambda / \lambda$; and similarly define $\bar{\boldsymbol{N}}^\lambda := \boldsymbol{N}^\lambda / \lambda$. We first establish the following functional weak law of large numbers (FWLLN) for $\bar{\boldsymbol{X}}$. The proof is given in Appendix A.2.

Theorem 1. (FWLLN in the QED regime.) *Suppose that the heavy-traffic QED assumptions (i)–(iii) are in force; and suppose that there exists a vector* $\boldsymbol{x}(0)$ *such that* $\bar{\boldsymbol{X}}^\lambda(0) \xrightarrow{\mathrm{D}} \boldsymbol{x}(0)$, *and*

$$\bar{\boldsymbol{P}}^\lambda(\cdot) := \boldsymbol{P}^\lambda(\lambda \cdot) \to \bar{\boldsymbol{P}}(\cdot) = [\bar{p}_{ij}(\cdot)]_{i,j=1}^K \tag{10}$$

as $\lambda \to \infty$, *where* $\bar{\boldsymbol{P}}(\cdot) \in \mathcal{M}^K$ *is some deterministic matrix-valued function, is Lipschitz, and has a spectral radius less than* 1*:* $\sup_{\zeta \in \mathbb{R}_+^K} r(\bar{\boldsymbol{P}}(\zeta)) < 1$. *Then, we have*

$$\bar{\boldsymbol{X}}^\lambda \xrightarrow{\mathrm{D}} \boldsymbol{x} \quad \text{in } (D^K, J_1) \text{ as } \lambda \to \infty, \tag{11}$$

where $\boldsymbol{x} = (x_1, \ldots, x_K)^\top$ *is the unique solution to the ODE*

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{a} - (I - \bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta(\boldsymbol{x}(t) - \boldsymbol{\nu})^+ - \Upsilon(\boldsymbol{x}(t) \wedge \boldsymbol{\nu}) \tag{12}$$

starting from $\boldsymbol{x}(0)$, *where* $\boldsymbol{a} = (a_1, \ldots, a_K)^\top$, $(\boldsymbol{x}(t) - \boldsymbol{\nu})^+ = ((x_1(t) - \nu_1)^+, (x_2(t) - \nu_2)^+, \ldots, (x_K(t) - \nu_K)^+)^\top$, $(\boldsymbol{x}(t) \wedge \boldsymbol{\nu}) = (x_1(t) \wedge \nu_1, x_2(t) \wedge \nu_2, \ldots, x_K(t) \wedge \nu_K)^\top$, $I = \mathrm{diag}\{1, \ldots, 1\}_{K \times K}$, $\Theta = \mathrm{diag}\{\theta_1, \ldots, \theta_K\}$, $\Upsilon = \mathrm{diag}\{\mu_1, \ldots, \mu_K\}$, *and* $\bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top$ *is the transpose of* $\bar{\boldsymbol{P}}(\boldsymbol{x}(t))$. *Moreover,* $\boldsymbol{x}(t) \to \boldsymbol{x}(\infty) := \boldsymbol{\nu}$ *as* $t \to \infty$.

The assumption on the spectral radius of $\bar{\boldsymbol{P}}(\cdot)$ guarantees that $I - \bar{\boldsymbol{P}}(\cdot)^\top$ is invertible; and when the switchover probabilities are constants (i.e. state independent), this assumption is equivalent to $\bar{\boldsymbol{P}}$ being substochastic, and, hence, $I - \bar{\boldsymbol{P}}^\top$ is an M-matrix; see Berman and Plemmons (1979). This is a standard assumption in the study of Jackson networks; see, e.g. Chen and Yao (2001). It ensures stability and the existence and uniqueness of the limiting fluid and diffusion processes. See also Corollary 3 below.

Direct verification confirms the steady-state solution to (12), $x_i(\infty) = \nu_i$ for each i. Thus, for large λ, $X_i^\lambda(\infty) \approx \nu_i \lambda \approx N_i^\lambda$, which implies that the queue length $Q_i^\lambda(\infty) = X_i^\lambda(\infty) - N_i^\lambda = o(\lambda)$. Moreover, it is interesting to note that the switchover probabilities p_{ij}, state dependent or not, do not play a role in the steady-state queue lengths.

Next, define the diffusion-scaled queue-length process $\hat{\boldsymbol{X}}^\lambda$ and virtual waiting-time process $\hat{\boldsymbol{V}}^\lambda$ by

$$\hat{\boldsymbol{X}}^\lambda := \sqrt{\lambda}(\bar{\boldsymbol{X}}^\lambda - \bar{\boldsymbol{N}}^\lambda) \quad \text{and} \quad \hat{\boldsymbol{V}}^\lambda := \sqrt{\lambda} \boldsymbol{V}^\lambda.$$

Note that here we center the process X_i^λ by its approximate steady-state value N_i^λ instead of its fluid limit $x_i(t)$. The functional central limit theorems (FCLTs) for $\hat{\boldsymbol{X}}^\lambda$ and $\hat{\boldsymbol{V}}^\lambda$ are presented in the following two theorems, each followed by some remarks, while their proofs are deferred to Appendix A.3.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

Theorem 2. (FCLT in the QED regime.) *Under the heavy-traffic QED assumptions (i)–(iii), if there exist a random vector* $\hat{\boldsymbol{X}}(0)$ *and a matrix-valued function* $\bar{\boldsymbol{P}}(\cdot) \in \mathcal{M}^K$ *such that* $\hat{\boldsymbol{X}}^\lambda(0) \xrightarrow{\mathrm{D}} \hat{\boldsymbol{X}}(0)$ *in* $\mathbb{R}^K$ *as* $\lambda \to \infty$, *and (10) holds with* $\bar{\boldsymbol{P}}$ *being assumed as in Theorem 1, then*

$$\hat{\boldsymbol{X}}^\lambda \xrightarrow{\mathrm{D}} \hat{\boldsymbol{X}} \quad \text{in } (D^K, J_1) \text{ as } \lambda \to \infty, \tag{13}$$

where $\hat{\boldsymbol{X}}$ *is the unique solution to the SDE*

$$\mathrm{d}\hat{\boldsymbol{X}}(t) = [\boldsymbol{\xi}(t) - \Upsilon\boldsymbol{\gamma} - (I - \bar{\boldsymbol{P}}(\boldsymbol{\nu})^\top)\Theta(\hat{\boldsymbol{X}}(t))^+ + \Upsilon(\hat{\boldsymbol{X}}(t))^-]\,\mathrm{d}t + \Sigma\,\mathrm{d}\boldsymbol{B}(t), \tag{14}$$

starting from $\hat{\boldsymbol{X}}(0)$, *where* $\boldsymbol{\xi}(t) = (\xi_1(t), \ldots, \xi_K(t))^\top$, $\gamma = (\gamma_1, \ldots, \gamma_K)^\top$, $\Sigma = \operatorname{diag}\{a_1 + \mu_1\nu_1, a_2 + \mu_2\nu_2, \ldots, a_K + \mu_K\nu_K\}^{1/2} = \sqrt{2}\operatorname{diag}\{a_1, \ldots, a_K\}^{1/2}$ *is the covariance coefficient, and* $\boldsymbol{B}$ *is a* K*-dimensional standard Brownian motion.*

Although the time-varying arrival rates do not affect the fluid limit $\boldsymbol{x}$ in (12), the effect on the diffusion limit $\hat{\boldsymbol{X}}$ is well represented by the term $\boldsymbol{\xi}(t)$ in (14). The switchover probabilities in the diffusion limit depend only on the steady state $\boldsymbol{\nu}$ of the fluid limit process $\boldsymbol{x}$. This is because, under the initial condition $\hat{\boldsymbol{X}}^\lambda(0) \xrightarrow{\mathrm{D}} \hat{\boldsymbol{X}}(0)$, the fluid-scaled process $\bar{\boldsymbol{X}}^\lambda(0) \xrightarrow{\mathrm{D}} \boldsymbol{\nu}$, so that the fluid limit $\boldsymbol{x}$ in (12) starts in the steady state and will remain there; hence, $\bar{\boldsymbol{X}}^\lambda \xrightarrow{\mathrm{D}} \boldsymbol{\nu}$ in D^K.

Theorem 3. (FCLT for the virtual waiting time process in the QED regime.) *Under the assumptions of Theorem 2,*

$$\hat{\boldsymbol{V}}^\lambda \xrightarrow{\mathrm{D}} \hat{\boldsymbol{V}} := (\tilde{\boldsymbol{\nu}}^{-1}\Upsilon^{-1})\hat{\boldsymbol{X}}^+ \quad \text{in } (D^K, J_1) \text{ as } \lambda \to \infty, \tag{15}$$

where $\hat{\boldsymbol{X}}$ *is defined in (14),* $\tilde{\boldsymbol{\nu}} = \operatorname{diag}\{\nu_1, \ldots, \nu_K\}$, *and the convergence is jointly with the limits in (13).*

The interaction among the service stations due to switchover is captured by the term $(I - \bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta(\boldsymbol{x}(t) - \boldsymbol{\nu})^+$ in the fluid limit $\boldsymbol{x}$ in (12), and by the term $(I - \bar{\boldsymbol{P}}(\boldsymbol{\nu})^\top)\Theta(\hat{\boldsymbol{X}}(t))^+$ in the diffusion limit $\hat{\boldsymbol{X}}$ in (14). Based on Theorems 2 and 3, we now obtain some comparison results for the queue lengths and virtual waiting times with respect to the switchover probabilities.

Corollary 1. *Consider two switchover probability matrices* $\boldsymbol{P}^{(1)}$ *and* $\boldsymbol{P}^{(2)}$ *such that*

$$\boldsymbol{P}^{(i),\lambda}(\lambda\cdot) \to \bar{\boldsymbol{P}}^{(i)}(\cdot) \quad \text{as } \lambda \to \infty,\ i = 1, 2,$$

with $\bar{\boldsymbol{P}}^{(1)}$ *and* $\bar{\boldsymbol{P}}^{(2)}$ *satisfying the assumptions in Theorem 1, and* $\bar{\boldsymbol{P}}^{(1)} \geq \bar{\boldsymbol{P}}^{(2)}$. *Suppose that the conditions in Theorem 2 hold, and let* $\hat{X}^{(i)}$ *be the diffusion limit in (14) with* $\bar{\boldsymbol{P}}^{(i)}$. *Then,* $\hat{X}^{(1)}(t) \geq \hat{X}^{(2)}(t)$ *for all* $t \geq 0$. *Moreover,*

$$(\hat{X}_i^{(1)}(t))^+ \geq (\hat{X}_i^{(2)}(t))^+ \quad \text{and} \quad (\hat{X}_i^{(1)}(t))^- \leq (\hat{X}_i^{(2)}(t))^- \quad \text{for all } t \geq 0, \tag{16}$$

for each $i, \ldots, K$, *and* $\hat{V}^{(1)}(t) \geq \hat{V}^{(2)}(t)$ *for all* $t \geq 0$.

Intuitively, the above comparison result says that increasing the switchover probabilities will allow more customers to stay inside the network, as opposed to abandonment. This will lead to longer queue lengths and waiting times, but will also result in improved server utilization and increased throughput. (Recall that $(\hat{X}_i(t))^-$ measures the number of idle servers.) This is consistent with the known fact that mobility enlarges the stability region in wireless networks; see Borst *et al.* (2006) and Grossglauser and Tse (2002).

When the arrival rates are constant, the stationary distribution π of the diffusion limit process $\hat{\boldsymbol{X}}(t)$ in (14) exists by a similar argument as in Dieker and Gao (2013) and can be characterized

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

by the basic adjoint relationship (BAR)

$$\int_{\mathbb{R}^K} Lf(x)\pi(x)\,\mathrm{d}x = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}^K),$$

where

$$Lf = \sum_{i=1}^{K}\left(2a_i\frac{\partial^2 f}{\partial x_i^2} + b_i(x)\frac{\partial f}{\partial x_i}\right),$$

a_i is given in (5), $b_i(x) = \sum_{k=1,k\neq i}^{K}\theta_k\bar{p}_{ki}(x)(x_k - \nu_k)^+ - \theta_i(1-\bar{p}_{ii}(x))(x_i-\nu_i)^+ - \mu(x_i\wedge\nu_i)$, and $C_b^2(\mathbb{R}^K)$ is the set of twice continuously differentiable functions f on $\mathbb{R}^K$ whose first- and second-order derivatives are bounded. We also refer the reader to Saure *et al*. (2009) and He and Dai (2013) for the computation of the stationary distribution.

For each t, we derive the mean and covariance functions of $\hat{\boldsymbol{X}}(t)$ in (14) as follows (for the covariance, apply Ito's formula; see, e.g. Karatzas and Shreve (1991)):

$$\begin{aligned}\mathrm{d}(\hat{\boldsymbol{X}}(t)\hat{\boldsymbol{X}}(t)^\top) &= \hat{\boldsymbol{X}}(t)(\mathrm{d}\hat{\boldsymbol{X}}(t)^\top) + (\mathrm{d}\hat{\boldsymbol{X}}(t))\hat{\boldsymbol{X}}(t)^\top + \mathrm{d}[\hat{\boldsymbol{X}}(t),\hat{\boldsymbol{X}}^\top(t)]\\ &= \hat{\boldsymbol{X}}(t)(\mathrm{d}\hat{\boldsymbol{X}}(t)^\top) + (\mathrm{d}\hat{\boldsymbol{X}}(t))\hat{\boldsymbol{X}}(t)^\top + \Sigma\Sigma^\top\,\mathrm{d}t.\end{aligned}$$

We remark that our results below coincide with Theorem 6.2 of Mandelbaum *et al*. (1998) in the case of constant switchover probabilities.

Corollary 2. *Let* $\boldsymbol{m}(t) = \mathbb{E}[\hat{\boldsymbol{X}}(t)]$ *and* $\boldsymbol{\Phi}(t) = \mathbb{E}[\hat{\boldsymbol{X}}(t)\hat{\boldsymbol{X}}(t)^\top]$ *for* $\hat{\boldsymbol{X}}$ *in (14). Then,* $\boldsymbol{m}(t)$ *and* $\boldsymbol{\Phi}(t)$ *satisfy the differential equations*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{m}(t) &= \boldsymbol{\xi}(t) - \Upsilon\boldsymbol{\gamma} - (I-\bar{\boldsymbol{P}}(\boldsymbol{\nu})^\top)\Theta\mathbb{E}[(\hat{\boldsymbol{X}}(t))^+] + \Upsilon\mathbb{E}[(\hat{\boldsymbol{X}}(t))^-],\\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Phi}(t) &= [\boldsymbol{\xi}(t) - \Upsilon\boldsymbol{\gamma}]\boldsymbol{m}(t)^\top + \boldsymbol{m}(t)[\boldsymbol{\xi}(t)^\top - \boldsymbol{\gamma}^\top\Upsilon^\top] - (I-\bar{\boldsymbol{P}}(\boldsymbol{\nu})^\top)\Theta\mathbb{E}[(\hat{\boldsymbol{X}}(t))^+\hat{\boldsymbol{X}}(t)^\top]\\ &\quad - \mathbb{E}[\hat{\boldsymbol{X}}(t)(\hat{\boldsymbol{X}}(t)^\top)^+]\Theta(I-\bar{\boldsymbol{P}}(\boldsymbol{\nu})) + \Upsilon\mathbb{E}[(\hat{\boldsymbol{X}}(t))^-\hat{\boldsymbol{X}}(t)^\top]\\ &\quad + \mathbb{E}[\hat{\boldsymbol{X}}(t)(\hat{\boldsymbol{X}}(t)^\top)^-]\Upsilon + \Sigma\Sigma^\top,\end{aligned}$$

with $\boldsymbol{m}(0) = \mathbb{E}[\hat{\boldsymbol{X}}(0)]$ *and* $\boldsymbol{\Phi}(0) = \mathbb{E}[\hat{\boldsymbol{X}}(0)\hat{\boldsymbol{X}}(0)^\top]$.

4. The ED regime

The ED regime is particularly relevant to networks with nonstationary arrivals. It is concerned with those transient time periods when the system is overloaded, and tries to capture the impact of time-dependent arrivals on system performance, queue lengths in particular.

Formally, the ED regime is characterized as follows. Similar to the scaling in the QED regime, we scale both the arrival rate and the number of servers by a (common) factor λ and write

$$\lambda_i(t) = \lambda a_i(t) \quad \text{and} \quad N_i^\lambda = \lambda\nu_i, \qquad i = 1,\dots,K, \tag{17}$$

where the positive constants ν_i and the nonnegative deterministic functions a_i satisfy, for any $T > 0$,

$$\inf_{0<t\le T}\left(\int_0^t a_i(s)\,\mathrm{d}s - \nu_i\mu_i t\right) > 0, \qquad i = 1,\dots,K. \tag{18}$$

In particular, the condition in (17) implies that the system is overloaded at all $t > 0$.

The following FWLLN holds for the fluid-scaled queue-length process $\bar{\boldsymbol{X}}^\lambda := \boldsymbol{X}^\lambda/\lambda$. The proof is presented in Appendix A.4.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

Theorem 4. (FWLLN in the ED regime.) *Suppose that the conditions in (17)–(18) hold; in addition, suppose that there exists a vector* $\boldsymbol{x}(0)$ *and a deterministic matrix-valued function* $\bar{\boldsymbol{P}}(\cdot)$ *in* $\mathcal{M}^K$ *such that* $\bar{\boldsymbol{X}}^\lambda(0) \xrightarrow{\mathrm{D}} \boldsymbol{x}(0)$, *and (10) holds as* $\lambda \to \infty$, *and* $\bar{\boldsymbol{P}}$ *satisfies the properties specified in Theorem 1. Then*

$$\bar{\boldsymbol{X}}^\lambda \xrightarrow{\mathrm{D}} \boldsymbol{x} \quad \text{in } (D^K, J_1) \text{ as } \lambda \to \infty,$$

where $\boldsymbol{x}$ *is the unique solution to the nonlinear ODE*

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{a}(t) - (I - \bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta(\boldsymbol{x}(t) - \boldsymbol{\nu})^+ - \Upsilon(\boldsymbol{x}(t) \wedge \boldsymbol{\nu}), \qquad t \geq 0, \tag{19}$$

starting from $\boldsymbol{x}(0)$, *where* $\boldsymbol{a}(t) = (a_1(t), \ldots, a_K(t))^\top$. *Moreover, there exists* $t_0 > 0$ *such that, for all* $t > t_0$, $\boldsymbol{x}(t) > \boldsymbol{\nu}$; *hence, (19) reduces to*

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{a}(t) - (I - \bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta(\boldsymbol{x}(t) - \boldsymbol{\nu}) - \Upsilon\boldsymbol{\nu}, \qquad t > t_0, \tag{20}$$

and if, in addition, the switchover probabilities $\bar{\boldsymbol{P}}(\cdot) = \boldsymbol{P}$ *are constant,*

$$\boldsymbol{x}(t) = \boldsymbol{x}(t_0) + \int_{t_0}^{t} \exp(-(I - \boldsymbol{P}^\top)\Theta(t-s))[\boldsymbol{a}(s) + ((I - \boldsymbol{P}^\top)\Theta - \Upsilon)\boldsymbol{\nu}]\,\mathrm{d}s, \qquad t \geq t_0. \tag{21}$$

The effect of time-varying arrival rates is manifest in the term $\boldsymbol{a}(t)$ on the right-hand side of (19). For instance, if $a_i(t)$ is a periodic function then so is $x(t)$. However, if $a_i(t)$ does converge to a constant as time evolves then the fluid limit $\boldsymbol{x}$ will have a constant steady state, as stated in the following corollary. If this steady-state result appears to be incompatible with the assumption in (17), which says that the input overpowers the service capacity at all time, one should bear in mind that the abandonment ensures the queue lengths will not grow to ∞.

Corollary 3. (Steady state of the fluid limit in the ED regime.) *Under the assumptions of Theorem 4, if there exists some positive constant vector* $\bar{\boldsymbol{a}}$ *such that* $\boldsymbol{a}(t) \to \bar{\boldsymbol{a}} > \Upsilon\boldsymbol{\nu}$ *as* $t \to \infty$ *then the solution* $\boldsymbol{x}(t)$ *to ODE (19) satisfies* $\boldsymbol{x}(t) \to \boldsymbol{x}(\infty) := \boldsymbol{x}^* = \boldsymbol{\nu} + \boldsymbol{q}^*$ *as* $t \to \infty$, *where* $\boldsymbol{q}^*$ *is the steady-state queue-length vector, solving the following nonlinear ODE:*

$$\bar{\boldsymbol{a}} - (I - \bar{\boldsymbol{P}}(\boldsymbol{\nu} + \boldsymbol{q}^*)^\top)\Theta\boldsymbol{q}^* - \Upsilon\boldsymbol{\nu} = 0. \tag{22}$$

If, in addition, the switchover probabilities $\bar{\boldsymbol{P}}$ *are constant then* $\boldsymbol{q}^*$ *has the explicit expression*

$$\boldsymbol{q}^* = \Theta^{-1}(I - \bar{\boldsymbol{P}}^\top)^{-1}(\bar{\boldsymbol{a}} - \Upsilon\boldsymbol{\nu}). \tag{23}$$

Next, define the diffusion-scaled processes $\hat{\boldsymbol{X}}^\lambda$ by

$$\hat{\boldsymbol{X}}^\lambda := \sqrt{\lambda}(\bar{\boldsymbol{X}}^\lambda - \boldsymbol{x}), \tag{24}$$

where $\boldsymbol{x}$ is defined in (19). Note that here the centering is carried out by the fluid limit $\boldsymbol{x}$ established in Theorem 4. The proof of the following FCLT limit is deferred to Appendix A.5.

Theorem 5. (FCLT in the ED regime.) *Suppose that the conditions in (17)–(18) hold; in addition, suppose that there exist random vectors* $\boldsymbol{x}(0)$ *and* $\hat{\boldsymbol{X}}(0)$, *and two deterministic matrix-valued functions* $\bar{\boldsymbol{P}}(\cdot)$, $\hat{\boldsymbol{P}}(\cdot)$ *in* $\mathcal{M}^K$ *such that* $\hat{\boldsymbol{X}}^\lambda(0) \xrightarrow{\mathrm{D}} \hat{\boldsymbol{X}}(0)$, $\boldsymbol{x}(0) > \boldsymbol{\nu}$, *(10) holds, and*

$$\hat{\boldsymbol{P}}^\lambda(\cdot) := \sqrt{\lambda}(\bar{\boldsymbol{P}}^\lambda(\cdot) - \bar{\boldsymbol{P}}(\cdot)) \to \hat{\boldsymbol{P}}(\cdot) = [\hat{p}_{ij}(\cdot)]_{i,j=1}^K \quad \text{as } \lambda \to \infty;$$

$\bar{\boldsymbol{P}}$ *and* $\hat{\boldsymbol{P}}$ *are Lipschitz, and the spectral radius of* $\bar{\boldsymbol{P}}$ *satisfies* $\sup_{\zeta \in \mathbb{R}^K} r(\bar{\boldsymbol{P}}(\zeta)) < 1$. *Then*

$$\hat{\boldsymbol{X}}^\lambda \xrightarrow{\mathrm{D}} \hat{\boldsymbol{X}} \quad \text{in } (D^K, J_1) \text{ as } \lambda \to \infty,$$

where $\hat{\boldsymbol{X}}$ *is the unique solution to the SDE*

$$\mathrm{d}\hat{\boldsymbol{X}}(t) = [\hat{\boldsymbol{P}}(\boldsymbol{x}(s))\Theta(\boldsymbol{x}(t) - \boldsymbol{\nu}) - (I - \bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta\hat{\boldsymbol{X}}(t)]\,\mathrm{d}t + \mathrm{d}\boldsymbol{W}(t), \qquad t \geq 0, \tag{25}$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

starting from $\hat{\boldsymbol{X}}(0)$*, where* $\boldsymbol{x}(t)$ *is defined in (20) starting with* $\boldsymbol{x}(0)$*, and* $\boldsymbol{W}$ *is a* K*-dimensional Brownian motion with mean* $\mathbf{0}$ *and covariance matrix* $\Sigma(t)\Sigma(t)^\top$ *defined by*

$$\Sigma(t)\Sigma(t)^\top := \operatorname{diag}\left\{\int_0^t \boldsymbol{a}(s)\,\mathrm{d}s + \Upsilon\boldsymbol{\nu}t + \int_0^t \tilde{\Psi}(\boldsymbol{x}(s))\Theta(\boldsymbol{x}(s)-\boldsymbol{\nu})\,\mathrm{d}s\right\} + \Delta(t), \tag{26}$$

where

$$\tilde{\Psi}(\cdot) = \begin{bmatrix} 1-\bar{p}_{11}(\cdot) & \bar{p}_{21}(\cdot) & \cdots & \bar{p}_{K1}(\cdot) \\ \bar{p}_{12}(\cdot) & 1-\bar{p}_{22}(\cdot) & \cdots & \bar{p}_{K2}(\cdot) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{p}_{1K}(\cdot) & \bar{p}_{2K}(\cdot) & \cdots & 1-\bar{p}_{KK}(\cdot) \end{bmatrix}, \tag{27}$$

$\Delta(t)$ *is a* $K \times K$ *matrix with* $\Delta_{ii}(t) = 0$ *for all* $i = 1, \ldots, K$*, and*

$$\Delta_{ij}(t) = -2\theta_i\theta_j\left(\int_0^t \bar{p}_{ij}(\boldsymbol{x}(s))(x_i(s)-\nu_i)\,\mathrm{d}s\right)\left(\int_0^t \bar{p}_{ji}(\boldsymbol{x}(s))(x_j(s)-\nu_j)\,\mathrm{d}s\right)$$

for all $i \neq j$*,* $i, j = 1, \ldots, K$*.*

Note that in the above theorem we assume that $\boldsymbol{x}(0) > \boldsymbol{\nu}$ so that the system starts in an overloaded state and the diffusion limit $\hat{\boldsymbol{X}}$ is valid starting from time 0. Otherwise, from Theorem 4, we know that there exists a time point $t_0 > 0$ such that $\boldsymbol{x}(t) > \boldsymbol{\nu}$ for all $t > t_0$; consequently, the diffusion limit applies starting from t_0.

The time-varying arrival rates affect both the drift term and the covariance coefficient matrix of the diffusion limit. In the drift term, the dependence on $\boldsymbol{a}(t)$ is indirect, through the fluid limit $\boldsymbol{x}$. In the covariance coefficient matrix, $\boldsymbol{a}(t)$ appears directly in the diagonal (the variance terms) and indirectly, also via $\boldsymbol{x}$, in the off-diagonal entries.

As in the fluid case, if $\boldsymbol{a}(t)$ is periodic then, clearly, $\hat{\boldsymbol{X}}$ will not reach a steady state. If $\boldsymbol{a}(t) \to \bar{\boldsymbol{a}}$ as $t \to \infty$, as in Corollary 3, and if (22) has a unique solution, then a unique steady state exists for the diffusion process $\hat{\boldsymbol{X}}$.

Corollary 4. (FCLT in the ED regime with constant switchover probability and constant arrival rates.) *Under the heavy-traffic ED assumptions (17)–(18), if there exists a random vector* $\hat{\boldsymbol{X}}(0)$ *such that* $\hat{\boldsymbol{X}}^\lambda(0) \xrightarrow{\mathrm{D}} \hat{\boldsymbol{X}}(0)$ *as* $\lambda \to \infty$*, and if the switchover probabilities and the arrival rates are all constant, then*

$$\hat{\boldsymbol{X}}^\lambda \xrightarrow{\mathrm{D}} \hat{\boldsymbol{X}} \quad \text{in } (D^K, J_1) \text{ as } \lambda \to \infty, \tag{28}$$

where $\hat{\boldsymbol{X}}$ *is the unique solution to the SDE*

$$\mathrm{d}\hat{\boldsymbol{X}}(t) = -(I - \boldsymbol{P}^\top)\Theta\hat{\boldsymbol{X}}(t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{W}(t), \qquad t \geq 0, \tag{29}$$

starting from $\hat{\boldsymbol{X}}(0)$*, where* $\boldsymbol{W}$ *is a* K*-dimensional Brownian motion with mean* $\mathbf{0}$ *and covariance matrix* $\Sigma\Sigma^\top$ *defined by*

$$\Sigma\Sigma^\top := \operatorname{diag}\{\boldsymbol{a} + \Upsilon\boldsymbol{\nu} + \tilde{\Psi}\Theta\boldsymbol{q}^*\} + \tilde{\Delta}, \tag{30}$$

where $\tilde{\Psi}$ *is defined in (27) with* $\bar{\boldsymbol{P}} = \boldsymbol{P}$ *being constant,* $\boldsymbol{q}^*$ *is given in (23), and* $\tilde{\Delta}$ *is defined by*

$$\tilde{\Delta} = \begin{bmatrix} 0 & -2\theta_1\theta_2 p_{12}p_{21}q_1^*q_2^* & \cdots & -2\theta_1\theta_K p_{1K}p_{K1}q_1^*q_K^* \\ -2\theta_1\theta_2 p_{12}p_{21}q_1^*q_2^* & 0 & \cdots & -2\theta_2\theta_K p_{2K}p_{K2}q_2^*q_K^* \\ \vdots & \vdots & \ddots & \vdots \\ -2\theta_1\theta_K p_{1K}p_{K1}q_1^*q_K^* & -2\theta_2\theta_K p_{2K}p_{K2}q_2^*q_K^* & \cdots & 0 \end{bmatrix}.$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

When the arrival rates and the switchover probabilities are constant, the diffusion limit simplifies substantially, which requires solving only SDE (29); however, with time-varying arrival rates and state-dependent switchover probabilities, the diffusion limit requires solving the coupled SDE (25) together with the ODE (20). Moreover, the covariance coefficient matrix becomes constant as well. With these constant parameters we can define the prelimit diffusion-scaled processes $\hat{\boldsymbol{X}}^\lambda$ by centering around the steady state of the fluid limit instead of centering around the fluid limit itself, which simplifies the analysis substantially; see the proof of Corollary 4.

Corollary 5. *Let* $\boldsymbol{m}(t) = \mathbb{E}[\hat{\boldsymbol{X}}(t)]$ *and* $\boldsymbol{\Phi}(t) = \mathbb{E}[\hat{\boldsymbol{X}}(t)\hat{\boldsymbol{X}}(t)^\top]$ *for* $\hat{\boldsymbol{X}}$ *in (25). Then,* $\boldsymbol{m}(t)$ *and* $\boldsymbol{\Phi}(t)$ *satisfy the differential equations*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{m}(t) &= \hat{\boldsymbol{P}}(\boldsymbol{x}(s))\Theta(\boldsymbol{x}(t)-\boldsymbol{\nu}) - (I-\bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta\boldsymbol{m}(t),\\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Phi}(t) &= \hat{\boldsymbol{P}}(\boldsymbol{x}(s))\Theta(\boldsymbol{x}(t)-\boldsymbol{\nu})\boldsymbol{m}(t)^\top - (I-\bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta\boldsymbol{\Phi}(t)\\
&\quad + \boldsymbol{m}(t)(\boldsymbol{x}(t)-\boldsymbol{\nu})^\top\Theta\hat{\boldsymbol{P}}(\boldsymbol{x}(s))^\top - \boldsymbol{\Phi}(t)\Theta(I-\bar{\boldsymbol{P}}(\boldsymbol{x}(t))) + \Sigma(t)\Sigma(t)^\top,
\end{aligned}$$

with $\boldsymbol{m}(0) = \mathbb{E}[\hat{\boldsymbol{X}}(0)]$ *and* $\boldsymbol{\Phi}(0) = \mathbb{E}[\hat{\boldsymbol{X}}(0)\hat{\boldsymbol{X}}(0)^\top]$, *where* $\boldsymbol{x}(t)$ *is defined in (20) and* $\Sigma(t)\Sigma(t)^\top$ *is defined in (26). When the switchover probabilities are constant with* $\bar{\boldsymbol{P}} = \boldsymbol{P}$ *and the arrival rates are time varying,*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{m}(t) &= -(I-\boldsymbol{P}^\top)\Theta\boldsymbol{m}(t),\\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Phi}(t) &= -(I-\boldsymbol{P}^\top)\Theta\boldsymbol{\Phi}(t) - \boldsymbol{\Phi}(t)\Theta(I-\boldsymbol{P}) + \Sigma(t)\Sigma(t)^\top,
\end{aligned} \tag{31}$$

where $\Sigma(t)\Sigma(t)^\top$ *is defined in (26) with* $\boldsymbol{x}(t)$ *equal to the explicit solution* ($t_0 = 0$) *in (21). If, in addition, the arrival rates are also constant then (31) holds with* $\Sigma(t)\Sigma(t)^\top$ *being replaced by* $\Sigma\Sigma^\top$ *defined in (30).*

5. Mixed regimes

In the last two sections, all stations in the network are operating either in the QED regime or in the ED regime. In many applications, it is often possible that some stations are overloaded while others are underloaded, and switchover will have the effect of redistributing the loads among the stations. In this section we consider mixed regimes in our multiclass network model. Suppose that the K stations are partitioned into three subgroups, $\mathcal{K}_1$, $\mathcal{K}_2$, and $\mathcal{K}_3$. For a station $i \in \mathcal{K}_1$, the corresponding parameters satisfy the QD assumption, that is, $a_i < \nu_i\mu_i$ for ν_i in (2) and a_i in (4). Similarly, for a station $i \in \mathcal{K}_2$, they satisfy the QED assumptions (2)–(8); and, for a station $i \in \mathcal{K}_3$, they satisfy the ED assumptions (17)–(18). We study the following mixed regimes:

QD/QED regime: $\mathcal{K}_1 \neq \varnothing$, $\mathcal{K}_2 \neq \varnothing$, $\mathcal{K}_3 = \varnothing$;

QED/ED regime: $\mathcal{K}_1 = \varnothing$, $\mathcal{K}_2 \neq \varnothing$, $\mathcal{K}_3 \neq \varnothing$;

QD/ED regime: $\mathcal{K}_1 \neq \varnothing$, $\mathcal{K}_2 = \varnothing$, $\mathcal{K}_3 \neq \varnothing$; and

QD/QED/ED regime: $\mathcal{K}_1 \neq \varnothing$, $\mathcal{K}_2 \neq \varnothing$, $\mathcal{K}_3 \neq \varnothing$.

We next discuss the fluid and diffusion limits of $\boldsymbol{X}^\lambda$ in these mixed regimes. For simplicity, we assume that the arrival rates are constant.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

Theorem 6. (FWLLN in the mixed regimes.) *For all four mixed regimes, under the assumptions on $\bar{\boldsymbol{X}}(0)$, $\bar{\boldsymbol{P}}^{\lambda}(\cdot)$, and $\bar{\boldsymbol{P}}(\cdot)$ in Theorem 1, the FWLLN in Theorem 1 holds for the fluid-scaled processes $\bar{\boldsymbol{X}}^{\lambda}$, that is, (11) holds with the limit $\boldsymbol{x}$ as the unique solution to the ODE in (12).*

The steady states $\boldsymbol{x}^*$ of the fluid limit $\boldsymbol{x}$ are as follows when the switchover probabilities $\bar{\boldsymbol{P}}$ are constant.

(i) *The mixed QD/QED regime: $x_i^* = a_i/\mu_i < \nu_i$ for $i \in \mathcal{K}_1$ and $x_i^* = a_i/\mu_i = \nu_i$ for $i \in \mathcal{K}_2$.*

(ii) *The mixed QED/ED regime: $x_i^* = \nu_i + q_i^*$ for $i \in \mathcal{K}_2 \cup \mathcal{K}_3$, where $q_i \geq 0$ is the ith component of $\boldsymbol{q}^* = \Theta^{-1}(I - \bar{\boldsymbol{P}}^{\top})^{-1}(\boldsymbol{a} - \Upsilon \cdot \boldsymbol{\nu})$, and $q_i^* = 0$ if $i \in \mathcal{K}_2$ and $\bar{p}_{ji} = 0$ for all $j \neq i$.*

(iii) *The mixed QD/ED regime: x^* is the unique solution to the set of equations*

$$a_i - \mu_i x_i^* + \sum_{j \in \tilde{\mathcal{K}}_3,\, j \neq i} \theta_j \bar{p}_{ji}(x_j^* - \nu_j) = 0 \quad \text{for } i \in \tilde{\mathcal{K}}_1, \tag{32}$$

$$a_i - \mu_i \nu_i - \theta_i(1 - \bar{p}_{ii})(x_i^* - \nu_i) + \sum_{j \in \tilde{\mathcal{K}}_3,\, j \neq i} \theta_j \bar{p}_{ji}(x_j^* - \nu_j) = 0 \quad \text{for } i \in \tilde{\mathcal{K}}_3, \tag{33}$$

$x_i^* = \nu_i$ for $i \in \tilde{\mathcal{K}}_2$, where $\tilde{\mathcal{K}}_1 = \{i : \tilde{q}_i^* < 0\} - \mathcal{K}_3$, $\tilde{\mathcal{K}}_2 = \{i : \tilde{q}_i^* = 0\}$, $\tilde{\mathcal{K}}_3 = \{i : \tilde{q}_i^* > 0\} \cup \mathcal{K}_3$, and $\tilde{\boldsymbol{q}}^* = \Theta^{-1}(I - \bar{\boldsymbol{P}}^{\top})^{-1}(\boldsymbol{a} - \Upsilon \boldsymbol{\nu})$.

(iv) *The mixed QD/QED/ED regime: x^* is the unique solution to the set of equations in (32).*

From the above theorem, we observe that in the mixed QD/QED regime, all queue lengths in steady state are 0 in the fluid scale; hence, the steady-state switchover quantities are 0. In the mixed QED/ED regime, as long as the switchover probability from the ED stations to a QED station is positive, the QED station will become overloaded in the steady state.

In the mixed QD/ED regime, the situation becomes more complex because a QD station can become overloaded if too much workload is switched over from the ED station. Consider an example of two stations to get a feel for the possible outcomes. Suppose that station 1 is ED and station 2 is QD; specifically, $a_1 > \mu_1 \nu_1$ and $a_2 < \mu_1 \nu_2$. We consider the following three cases.

Case 1: $\tilde{q}_1^* > 0$ *and* $\tilde{q}_2^* < 0$. That is, $(1 - \bar{p}_{22})(a_1 - \mu_1 \nu_1) > \bar{p}_{21}(\mu_2 \nu_2 - a_2)$ and $\bar{p}_{12}(a_1 - \mu_1 \nu_1) < (1 - \bar{p}_{11})(\mu_2 \nu_2 - a_2)$. Then $\boldsymbol{x}^* = (x_1^*, x_2^*)^{\top}$ solves the equations

$$a_1 - \mu_1 \nu_1 - \theta_1(1 - \bar{p}_{11})(x_1^* - \nu_1) = 0, \qquad a_2 + \theta_1 \bar{p}_{12}(x_1^* - \nu_1) - \mu_2 x_2^* = 0,$$

and

$$x_1^* = \nu_1 + \frac{a_1 - \mu_1 \nu_1}{\theta_1(1 - \bar{p}_{11})} > \nu_1, \qquad x_2^* = \frac{a_2}{\mu_2} + \frac{\bar{p}_{12}(a_1 - \mu_1 \nu_1)}{\mu_2(1 - \bar{p}_{11})} < \nu_2.$$

Thus, in this case, station 1 remains overloaded and station 2 remains underloaded. The amount of work switched from station 1 over to station 2 is equal to $x_2^* - a_2/\mu_2$.

Case 2: $\tilde{q}_1^* > 0$ *and* $\tilde{q}_2^* > 0$. That is, $(1 - \bar{p}_{22})(a_1 - \mu_1 \nu_1) > \bar{p}_{21}(\mu_2 \nu_2 - a_2)$ and $\bar{p}_{12}(a_1 - \mu_1 \nu_1) > (1 - \bar{p}_{11})(\mu_2 \nu_2 - a_2)$. Then $\boldsymbol{x}^* = (x_1^*, x_2^*)^{\top}$ solves the equations

$$a_1 - \mu_1 \nu_1 - \theta_1(1 - \bar{p}_{11})(x_1^* - \nu_1) + \theta_2 \bar{p}_{21}(x_2^* - \nu_2) = 0,$$
$$a_2 - \mu_2 \nu_2 - \theta_2(1 - \bar{p}_{22})(x_2^* - \nu_2) + \theta_1 \bar{p}_{12}(x_1^* - \nu_1) = 0,$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

and

$$x_1^* = \nu_1 + \tilde{q}_1^* = \nu_1 + \frac{(1-\bar{p}_{22})(a_1-\mu_1\nu_1) + \bar{p}_{21}(a_2-\mu_2\nu_2)}{\theta_1((1-\bar{p}_{11})(1-\bar{p}_{22}) - \bar{p}_{12}\bar{p}_{21})} > \nu_1,$$
$$x_2^* = \nu_2 + \tilde{q}_2^* = \nu_2 + \frac{\bar{p}_{12}(a_1-\mu_1\nu_1) + (1-\bar{p}_{11})(a_2-\mu_2\nu_2)}{\theta_2((1-\bar{p}_{11})(1-\bar{p}_{22}) - \bar{p}_{12}\bar{p}_{21})} > \nu_2.$$

Thus, in this case, both stations are overloaded. The amount of work switched over from station 1 to station 2 is equal to $x_2^* - a_2/\mu_2$.

Case 3: $\tilde{q}_1^* > 0$ *and* $\tilde{q}_2^* = 0$. That is, $(1-\bar{p}_{22})(a_1-\mu_1\nu_1) > \bar{p}_{21}(\mu_2\nu_2 - a_2)$ and $\bar{p}_{12}(a_1 - \mu_1\nu_1) = (1-\bar{p}_{11})(\mu_2\nu_2 - a_2)$. Then $\boldsymbol{x}^* = (x_1^*, x_2^*)^\top$ is a solution to the same equations as in the case when $\tilde{q}_1^* > 0$ and $\tilde{q}_2^* < 0$:

$$x_1^* = \nu_1 + \frac{a_1 - \mu_1\nu_1}{\theta_1(1-\bar{p}_{11})} > \nu_1, \qquad x_2^* = \frac{a_2}{\mu_2} + \frac{\bar{p}_{12}(a_1-\mu_1\nu_1)}{\mu_2(1-\bar{p}_{11})} = \nu_2.$$

Thus, in this case, station 1 remains overloaded, while station 2 becomes critically loaded, and the switchover amount is equal to $x_2^* - a_2/\mu_2 = \nu_2 - a_2/\mu_2$. Moreover, it is easy to check from the fluid equation of $\boldsymbol{x}(t)$ in (12) that it is impossible to have $x_1^* < \nu_1$ and $x_2^* \geq \nu_2$. In the mixed QD/QED/ED regime, we observe the same phenomena as in the QD/ED regime.

We now present the diffusion limits in the four mixed regimes.

Theorem 7. (FCLT in the mixed regimes.) *Suppose that the assumptions on the transitions probabilities in Theorem 6 hold.*

(i) *The mixed QD/QED regime. Define the diffusion-scaled processes by* $\hat{X}_i^\lambda := \sqrt{\lambda}(\bar{X}_i^\lambda - \lambda_i/\mu_i)$ *for* $i \in \mathcal{K}_1$ *and* $\hat{X}_i^\lambda := \sqrt{\lambda}(\bar{X}_i^\lambda - \bar{N}_i^\lambda)$ *for* $i \in \mathcal{K}_2$. *If there exist random variables* $\hat{X}_i(0)$ *such that* $\hat{X}_i^\lambda(0) \xrightarrow{\mathrm{D}} \hat{X}_i(0)$ *as* $\lambda \to \infty$, *then (13) holds with the limit* $\hat{X}_i$ *being the unique solution to the stochastic integral equation*

$$\begin{aligned}
\hat{X}_i(t) &= \hat{X}_i(0) + \int_0^t \Bigg(\sum_{k\in\mathcal{K}_2} \theta_k \bar{p}_{ki} (\hat{X}_k(s))^+ \Bigg) \mathrm{d}s - \mu_i \int_0^t \hat{X}_i(s)\,\mathrm{d}s \\
&\quad + \sqrt{2a_i} B_i(t) \quad \text{for } i \in \mathcal{K}_1, \qquad (34)\\
\hat{X}_i(t) &= \hat{X}_i(0) + (\xi_i - \mu_i\gamma_i)t \\
&\quad + \int_0^t \Bigg(\sum_{k\in\mathcal{K}_2,\, k\neq i} \theta_k \bar{p}_{ki} (\hat{X}_k(s))^+ - \theta_i(1-\bar{p}_{ii})(\hat{X}_i(s))^+ \Bigg) \mathrm{d}s \\
&\quad + \mu_i \int_0^t (\hat{X}_i(s))^- \,\mathrm{d}s + \sqrt{2a_i} B_i(t) \quad \text{for } i \in \mathcal{K}_2,
\end{aligned}$$

where $\boldsymbol{B} = (B_1, \ldots, B_K)$ *is a* K*-dimensional standard Brownian motion.*

(ii) *The mixed QED/ED regime. Here the diffusion-scaled process* $\hat{\boldsymbol{X}}^\lambda$ *is the same as in (24). If there exist random variables* $\hat{X}_i(0)$ *such that* $\hat{X}_i^\lambda(0) \xrightarrow{\mathrm{D}} \hat{X}_i(0)$ *as* $\lambda \to \infty$, *the FCLT in Corollary 4 holds.*

(iii) *The mixed QD/ED regime. Define the diffusion-scaled process by* $\hat{X}_i^\lambda := \sqrt{\lambda}(\bar{X}_i^\lambda - x_i^*)$, *where* x_i^* *is defined in (32). If there exist random variables* $\hat{X}_i(0)$ *such that* $\hat{X}_i^\lambda(0) \xrightarrow{\mathrm{D}} \hat{X}_i(0)$ *as* $\lambda \to \infty$ *then (13) holds with the limit* $\hat{X}_i$ *being the unique solution to the stochastic*

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

integral equation

$$\hat{X}_i(t) = \hat{X}_i(0) + \int_0^t \Bigg(\sum_{k\in\tilde{\mathcal{K}}_2} \theta_k \bar{p}_{ki} (\hat{X}_k(s))^+ + \sum_{k\in\tilde{\mathcal{K}}_3} \theta_k \bar{p}_{ki} \hat{X}_k(s) \Bigg) \mathrm{d}s - \mu_i \int_0^t \hat{X}_i(s)\,\mathrm{d}s + W_i(t) \quad \text{for } i \in \tilde{\mathcal{K}}_1, \tag{35}$$

$$\hat{X}_i(t) = \hat{X}_i(0) + \int_0^t \Bigg(\sum_{k\in\tilde{\mathcal{K}}_2,\, k\neq i} \theta_k \bar{p}_{ki} (\hat{X}_k(s))^+ + \sum_{k\in\tilde{\mathcal{K}}_3} \theta_k \bar{p}_{ki} \hat{X}_k(s) \Bigg) \mathrm{d}s - \int_0^t (\theta_i(1-\bar{p}_{ii})(\hat{X}_i(s))^+)\,\mathrm{d}s + \mu_i \int_0^t (\hat{X}_i(s))^-\,\mathrm{d}s + W_i(t) \quad \text{for } i \in \tilde{\mathcal{K}}_2, \tag{36}$$

$$\hat{X}_i(t) = \hat{X}_i(0) + \int_0^t \Bigg(\sum_{k\in\tilde{\mathcal{K}}_2} \theta_k \bar{p}_{ki} (\hat{X}_k(s))^+ + \sum_{k\in\tilde{\mathcal{K}}_3,\, k\neq i} \theta_k \bar{p}_{ki} \hat{X}_k(s) \Bigg) \mathrm{d}s - \int_0^t (\theta_i(1-\bar{p}_{ii})\hat{X}_i(s))\,\mathrm{d}s + W_i(t) \quad \text{for } i \in \tilde{\mathcal{K}}_3, \tag{37}$$

where $\boldsymbol{W} = (W_1, \ldots, W_K)^\top$ *is a* K*-dimensional Brownian motion with covariance matrix* $[\sigma_{ij}]_{i,j=1,\ldots,K}$ *given by*

$$\begin{aligned}
\sigma_{ii} &= a_i + \mu_i x_i^* + \sum_{k\in\tilde{\mathcal{K}}_3} \theta_k p_{ki}(x_k^* - \nu_k), \qquad i \in \tilde{\mathcal{K}}_1,\\
\sigma_{ii} &= 2a_i + \sum_{k\in\tilde{\mathcal{K}}_3} \theta_k p_{ki}(x_k^* - \nu_k), \qquad i \in \tilde{\mathcal{K}}_2,\\
\sigma_{ii} &= a_i + \mu_i \nu_i^* + \sum_{k=1,\, k\neq i}^{K} \theta_i p_{ik}(x_k^* - \nu_k) + \theta_i(1-p_{ii})(x_i^* - \nu_i)\\
&\quad + \sum_{k\in\tilde{\mathcal{K}}_3,\, k\neq i} \theta_k p_{ki}(x_k^* - \nu_k), \qquad i \in \tilde{\mathcal{K}}_3,\\
\sigma_{ij} &= \sigma_{ji} = 0, \qquad i \in \tilde{\mathcal{K}}_1,\ j \in \tilde{\mathcal{K}}_2 \cup \tilde{\mathcal{K}}_3 \text{ or } i \in \tilde{\mathcal{K}}_1 \cup \tilde{\mathcal{K}}_2,\ j \in \tilde{\mathcal{K}}_3,\\
\sigma_{ij} &= \sigma_{ji} = -2\theta_i\theta_j \bar{p}_{ij}\bar{p}_{ji}(x_i^* - \nu_i)(x_j^* - \nu_j), \qquad i \in \tilde{\mathcal{K}}_3,\ j \in \tilde{\mathcal{K}}_3.
\end{aligned}$$

(iv) *The mixed QD/QED/ED regime. The FCLT for* $\hat{\boldsymbol{X}}^\lambda$ *holds with the same limiting diffusion processes as in case (iii).*

We remark that in the mixed QD/QED regime, switchover has an effect on the diffusion limit even though it does not on the fluid limit. In particular, there is an extra term in the drift term of the diffusion limit $\hat{X}_i$ for QD stations, which captures the impact of the switchover from the QED stations; see the second term in (34). The same happens in the mixed QD/ED regime and in the mixed QD/QED/ED regime; see the second term of the diffusion limit $\hat{X}_i$ in (35). In addition, the QED stations also have switchover impact upon the ED regimes; see the second term of the diffusion limit $\hat{X}_i$ in (37).

6. Concluding remarks

As motivated in the introduction, the model we have studied in this paper, a network of many-server stations allowing waiting jobs to switch across stations, along with other features such as nonstationary arrivals and abandonment, has wide-ranging applications, including healthcare

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

delivery and mobile communication. These applications typically support a high volume, time-varying demand pattern and involve expensive resources that are, quite naturally, heavily utilized. We have analyzed not only the QED and ED regimes among all stations, but also the mixture of these regimes and the QD regime, in any combination. These mixed regimes capture the behavior of the network in the presence of supply-demand imbalance, due to demand (traffic) volatility or nonstationarity, whereas the switchover mechanism migrates the workload from overloaded stations to less loaded stations, achieving a load-balancing effect. Key performance measures under all regimes are derived in the form of fluid and diffusion limits, and characterized by differential equations (ODE or SDE). The latter, in turn, provides numerical means to estimate the performance measures. They also provide the dynamics of queue-length and workload processes, should one pursue an optimization or optimal control on some operational aspects of the network.

Appendix A. Proofs

A.1. A martingale representation of the queueing processes

In this subsection we present and prove a martingale representation of the queue-length process $\boldsymbol{X}$ in (1). The fluid-scaled and diffusion-scaled queue-length processes $\bar{\boldsymbol{X}}^\lambda$ and $\hat{\boldsymbol{X}}^\lambda$ in the QED and ED regimes also have the same type of martingale representations with corresponding fluid-scaled and diffusion-scaled martingale terms, which will be stated in the proofs of the FWLLN and FCLT theorems in the following sections. The proof follows a similar argument as the proof of Theorem 7.2 of Pang *et al*. (2007) and is thus omitted.

Lemma 1. (Martingale representation of queue-length processes.) *The K-dimensional queue-length process $\boldsymbol{X}$ in (1) has the martingale representation*

$$\begin{aligned}\boldsymbol{X}(t) = \boldsymbol{X}(0) + \Lambda(t) - \int_0^t (I - \boldsymbol{P}(\boldsymbol{X}(s-))^\top)\Theta(\boldsymbol{X}(s) - \boldsymbol{N})^+ \,\mathrm{d}s - \int_0^t \Upsilon(\boldsymbol{X}(s) \wedge \boldsymbol{N})\,\mathrm{d}s \\ + \boldsymbol{M}_A(t) - \boldsymbol{M}_{L,0}(t) - \boldsymbol{M}_S(t) + \sum_{i=1}^K [\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\boldsymbol{M}_L(t) - \boldsymbol{M}_L(t)\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1} \qquad (38)\end{aligned}$$

for each $t \geq 0$, where $\Lambda(t) := (\int_0^t \lambda_1(s)\,\mathrm{d}s, \ldots, \int_0^t \lambda_K(s)\,\mathrm{d}s)^\top$, $\boldsymbol{N} := (N_1, \ldots, N_K)^\top$, $(\boldsymbol{X}(s) - \boldsymbol{N})^+ := ((X_1(s) - N_1)^+, \ldots, (X_K(s) - N_K)^+)^\top$, $\boldsymbol{X}(s) \wedge \boldsymbol{N} := (X_1(s) \wedge N_1, \ldots, X_K(s) \wedge N_K)^\top$, $\boldsymbol{P}(\boldsymbol{X}(t))^\top$ is the transpose of $\boldsymbol{P}(\boldsymbol{X}(t))$, $\tilde{\boldsymbol{E}} = \boldsymbol{E} - I$ with $\boldsymbol{E}$ a $K \times K$ matrix whose components equal 1, $\boldsymbol{E}^{(ii)}$ is a $K \times K$ matrix whose (i,i)th component is 1 and whose other components are 0, $\mathbf{1}$ is a K-dimensional vector of 1s, $\boldsymbol{M}_A(t) = (M_{A,1}(t), \ldots, M_{A,K}(t))^\top$, $\boldsymbol{M}_S(t) = (M_{S,1}(t), \ldots, M_{S,K}(t))^\top$, and $\boldsymbol{M}_{L,0}(t) = (M_{L,1,0}(t), \ldots, M_{L,K,0}(t))^\top$ are all K-dimensional square-integrable $\boldsymbol{F}$-martingales, $\boldsymbol{M}_L(t) = [M_{L,i,j}(t)]_{i,j=1}^K$ is a $(K \times K)$-dimensional square-integrable $\boldsymbol{F}$-martingale matrix, the filtration $\boldsymbol{F} := \{\mathcal{F}(t) : t \geq 0\}$ is defined by

$$\begin{aligned}\mathcal{F}(t) = \sigma\Bigg\{ X_i(0), L_{i,j}\bigg(\theta_i \int_0^s p_{ij}(\boldsymbol{X}(s-))(X_i(u) - N_i)^+ \,\mathrm{d}s\bigg), \\ S_i\bigg(\mu_i \int_0^s (X_i(u) \wedge N_i)\,\mathrm{d}u\bigg) : 0 \leq s \leq t,\ i = 1, \ldots, K,\ j = 0, 1, \ldots, K \Bigg\} \\ \vee \sigma\Bigg\{ A_i\bigg(\int_0^t \lambda_i(s)\,\mathrm{d}s\bigg) : t \geq 0,\ i = 1, \ldots, K \Bigg\} \vee \mathcal{N}\end{aligned}$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

with $\mathcal{N}$ being the collection of all null sets, and

$$M_{A,i}(t) = A_i\left(\int_0^t \lambda_i(s)\,\mathrm{d}s\right) - \int_0^t \lambda_i(s)\,\mathrm{d}s,$$
$$M_{S,i}(t) = S_i\left(\mu_i \int_0^t (X_i(s) \wedge N_i)\,\mathrm{d}s\right) - \mu_i \int_0^t (X_i(s) \wedge N_i)\,\mathrm{d}s,$$
$$M_{L,i,j}(t) = L_{i,j}\left(\theta_i \int_0^t p_{ij}(\boldsymbol{X}(s-))(X_i(s) - N_i)^+\,\mathrm{d}s\right) - \theta_i \int_0^t p_{ij}(\boldsymbol{X}(s-))(X_i(s) - N_i)^+\,\mathrm{d}s,$$

and their predictable quadratic variations are given by

$$\langle M_{A,i}\rangle(t) = \int_0^t \lambda_i(s)\,\mathrm{d}s, \qquad \langle M_{S,i}\rangle(t) = \mu_i \int_0^t (X_i(s) \wedge N_i)\,\mathrm{d}s,$$
$$\langle M_{L,i,j}\rangle(t) = \theta_i \int_0^t p_{ij}(\boldsymbol{X}(s-))(X_i(s) - N_i)^+\,\mathrm{d}s.$$

A.2. Proof of the fluid limit in the QED regime

First, by (38), we have the following martingale representation of the fluid-scaled queue-length process $\bar{\boldsymbol{X}}^\lambda$:

$$\bar{\boldsymbol{X}}^\lambda(t) = \bar{\boldsymbol{X}}^\lambda(0) + \bar{\Lambda}^\lambda(t) - \int_0^t [(I - \boldsymbol{P}^\lambda(\lambda\bar{\boldsymbol{X}}^\lambda(s-))^\top)\Theta(\bar{\boldsymbol{X}}^\lambda(s) - \bar{\boldsymbol{N}}^\lambda)^+ + \Upsilon(\bar{\boldsymbol{X}}^\lambda(s) \wedge \bar{\boldsymbol{N}}^\lambda)]\,\mathrm{d}s + \bar{\boldsymbol{M}}_A^\lambda(t) - \bar{\boldsymbol{M}}_{L,0}^\lambda(t) - \bar{\boldsymbol{M}}_S^\lambda(t) + \sum_{i=1}^K [\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\bar{\boldsymbol{M}}_L(t) - \bar{\boldsymbol{M}}_L(t)\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1}. \quad (39)$$

Here

$$\bar{\Lambda}^\lambda(t) = \frac{1}{\lambda}\Lambda^\lambda(t) = \frac{1}{\lambda}\left(\int_0^t \lambda_1(s)\,\mathrm{d}s, \ldots, \int_0^t \lambda_K(s)\,\mathrm{d}s\right)^\top,$$

$\bar{\boldsymbol{M}}_A^\lambda(t) = \lambda^{-1}\boldsymbol{M}_A(t)$ and similarly for $\bar{\boldsymbol{M}}_L^\lambda(t)$, $\bar{\boldsymbol{M}}_{L,0}^\lambda(t)$, and $\bar{\boldsymbol{M}}_S^\lambda(t)$.

We will follow the procedure of the martingale proof approach reviewed in Pang *et al.* (2007). The major difference is that we have to prove convergence of multidimensional processes, instead of one-dimensional processes. We will again apply the CMT, but the mapping is defined through a multidimensional integral representation, as in (39). The continuity of such a mapping in the Skorokhod topology is not an easy generalization of that in the one-dimensional case.

Lemma 2. (Continuity of a multidimensional integral representation in the Skorokhod J_1 topology.) *Consider the K-dimensional integral representation*

$$x(t) = b + y(t) - \int_0^t [(I - P(x(s))^\top)\Theta(x(s) - a)^+ + \Upsilon(x(s) \wedge a)]\,\mathrm{d}s, \qquad t \geq 0, \quad (40)$$

where $a, b \in \mathbb{R}^K$, $y \in D^K$, and $\mathcal{M}^K \ni P(\cdot)\colon \mathbb{R}^K \to [0,1]^{K\times K}$ is Lipschitz, that is, $\|P(z_1) - P(z_2)\| \leq c_P\|z_1 - z_2\|$ for some positive constant c_P and for any $z_1, z_2 \in \mathbb{R}^K$, $P^\top$ is the

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

transpose of P. Then this integral representation has a unique solution, which defines a mapping $\phi\colon D^K \times \mathbb{R}^K \times \mathbb{R}^K \times \mathcal{M}^K \to D^K$ that maps (y, a, b, P) into $x := \phi(y, a, b, P)$. Moreover, the mapping ϕ is continuous in the space D^K endowed with the topology of uniform convergence over bounded intervals, or the Skorokhod J_1 topology. If $y \in C^K$ then $x \in C^K$.

Proof. The proof of existence and uniqueness is straightforward and thus omitted. In fact, the uniqueness argument follows easily from the proof of the continuity in the topology of uniform convergence over bounded intervals. Since the proof of the continuity in this topology is similar and easier than that in the Skorokhod topology, we only state the proof of the latter. We want to show that $x_n \to x$ in (D^K, J_1) when $(a_n, b_n, P_n, y_n) \to (a, b, P, y)$ in $\mathbb{R}^K \times \mathbb{R}^K \times \mathcal{M}^K \times D^K$ as $n \to \infty$.

Fix T to be a continuity point of y. By the convergence of $y_n \to y$ in (D^K, J_1), there exist increasing homeomorphisms κ_n of the interval $[0, T]$ such that $\|y_n - y \circ \kappa_n\|_T \to 0$ and $\|\kappa_n - e\|_T \to 0$ as $n \to \infty$, where $e(t) = t$ for each $t \geq 0$. Moreover, we can choose homeomorphisms κ_n to be absolutely continuous with respect to the Lebesgue measure on $[0, T]$ such that their derivatives $\dot{\kappa}_n$ satisfy $\|\dot{\kappa}_n - 1\|_T \to 0$ as $n \to \infty$. We again use the fact that functions in D^K are bounded so that, given a, b, P, and y, $x = \phi(a, b, P, y)$ is in the space D^K, and we let $M_x := \sup_{0\leq t\leq T} \|x(t)\| < \infty$. Then, we have

$$
\begin{aligned}
&\|x_n(t) - x(\kappa_n(t))\| \\
&\quad \leq \|b_n - b\| + \|y_n - y \circ \kappa_n\|_T \\
&\qquad + \left\| \int_0^t [(I - P_n(x_n(u))^\top)\Theta(x_n(u) - a_n)^+ + \Upsilon(x_n(u) \wedge a_n)]\,\mathrm{d}u \right. \\
&\qquad\quad \left. - \int_0^t \dot{\kappa}_n(u)[(I - P(x(\kappa_n(u)))^\top)\Theta(x(\kappa_n(u)) - a)^+ + \Upsilon(x(\kappa_n(u)) \wedge a)]\,\mathrm{d}u \right\| \\
&\quad \leq \|b_n - b\| + \|y_n - y \circ \kappa_n\|_T + \|\Upsilon\|\|a_n - a\|T \\
&\qquad + 2((K^2\|\Theta\|) \vee \|\Upsilon\|)\left(\|a\|T + \int_0^T \|x(s)\|\,\mathrm{d}s\right)\|\dot{\kappa}_n - 1\|_T \\
&\qquad + \left\| \int_0^t [(I - P_n(x_n(u))^\top)\Theta(x_n(u) - a_n)^+ - \Upsilon(x_n(u) - a_n)^- \right. \\
&\qquad\qquad \left. - (I - P_n(x_n(u))^\top)\Theta(x(\kappa_n(u)) - a)^+ + \Upsilon(x(\kappa_n(u)) - a)^-]\,\mathrm{d}u \right\| \\
&\qquad + \left\| \int_0^t [(P_n(x_n(u))^\top - P(x_n(u))^\top)\Theta(x(\kappa_n(u)) - a)^+]\,\mathrm{d}u \right\| \\
&\qquad + \left\| \int_0^t [(P(x_n(u))^\top - P(x(\kappa_n(u)))^\top)\Theta(x(\kappa_n(u)) - a)^+]\,\mathrm{d}u \right\| \\
&\quad \leq \|b_n - b\| + \|y_n - y \circ \kappa_n\|_T + \|\Upsilon\|\|a_n - a\|T \\
&\qquad + 2((K^2\|\Theta\|) \vee \|\Upsilon\|)\left(\|a\|T + \int_0^T \|x(s)\|\,\mathrm{d}s\right)\|\dot{\kappa}_n - 1\|_T \\
&\qquad + (K^2\|\Theta\| \vee \|\Upsilon\|)\int_0^T \|(x_n(u) - a_n) - (x(\kappa_n(u)) - a)\|\,\mathrm{d}u \\
&\qquad + \|\Theta\|(M_x + \|a\|)\int_0^T (\|P_n(x_n(u))^\top - P(x_n(u)))^\top\| + c_P\|x_n(u) - x(\kappa_n(u))\|)\,\mathrm{d}u
\end{aligned}
$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

$$
\begin{aligned}
&\le \|b_n - b\| + \|y_n - y \circ \kappa_n\|_T + \|\Upsilon\|\|a_n - a\|T \\
&\quad + 2((K^2\|\Theta\|) \vee \|\Upsilon\|)\left(\|a\|T + \int_0^T \|x(s)\|\,\mathrm{d}s\right)\|\dot{\kappa}_n - 1\|_T \\
&\quad + (K^2\|\Theta\| \vee \|\Upsilon\|)\int_0^T \|x_n(u) - x(\kappa_n(u))\|\mathrm{d}u + (K^2\|\Theta\| \vee \|\Upsilon\|)\|a_n - a\|T \\
&\quad + \|\Theta\|(M_x + \|a\|)\int_0^T (\|P_n(x_n(u))^\top - P(x_n(u))^\top\| + c_P\|x_n(u) - x(\kappa_n(u))\|)\,\mathrm{d}u \\
&\le \|b_n - b\| + \|y_n - y \circ \kappa_n\|_T + \|a_n - a\|(\|\Upsilon\| + K^2\|\Theta\| \vee \|\Upsilon\|)T \\
&\quad + 2((K^2\|\Theta\|) \vee \|\Upsilon\|)(\|a\| + M_x)T\|\dot{\kappa}_n - 1\|_T \\
&\quad + \|\Theta\|(M_x + \|a\|)\int_0^T \|P_n(x_n(u))^\top - P(x_n(u))^\top\|\,\mathrm{d}u \\
&\quad + [K^2\|\Theta\| \vee \|\Upsilon\| + c_P\|\Theta\|(M_x + \|a\|)]\int_0^T \|x_n(u) - x(\kappa_n(u))\|\,\mathrm{d}u.
\end{aligned}
$$

Choose n_0 large enough such that

$$
\|b_n - b\| < \frac{\delta}{5}, \qquad \|a_n - a\| < \frac{\delta}{5(\|\Upsilon\| + K^2\|\Theta\| \vee \|\Upsilon\|)T},
$$
$$
\|y_n - y \circ \kappa_n\|_T < \frac{\delta}{5}, \qquad \|\dot{\kappa}_n - 1\|_T < \frac{\delta}{10((K^2\|\Theta\|) \vee \|\Upsilon\|)(\|a\| + M_x)T},
$$

and

$$
\|P_n - P\| \le \frac{\delta}{5\|\Theta\|(M_x + \|a\|)T}.
$$

Let $\tilde{c} := [K^2\|\Theta\| \vee \|\Upsilon\| + c_P\|\Theta\|(M_x + \|a\|)]$. Then, by Gronwall's inequality,

$$
\|x_n(t) - x(\kappa_n(t))\| \le \delta \mathrm{e}^{\tilde{c}T}, \qquad 0 \le t \le T.
$$

Thus, $\|x_n - x \circ \kappa_n\|_T \le \delta \mathrm{e}^{\tilde{c}T}$. Finally, for any given $\varepsilon > 0$, we can choose δ small such that $\|x_n - x \circ \kappa_n\|_T \le \varepsilon$ for all $n \ge n_0$. This completes the proof.

We also prove the preservation of SB of a multidimensional integral representation, which generalizes the one-dimensional case, Lemma 5.5 of Pang *et al*. (2007).

Lemma 3. (SB of a multidimensional integral representation.) *Consider the K-dimensional integral representation of stochastic processes*

$$
\begin{aligned}
X_n(t) &= X_n(0) + Y_{n,1}(t) + \cdots + Y_{n,K}(t) \\
&\quad - \int_0^t [(I - P_n(X_n(s))^\top)\Theta(X_n(s) - Z_n)^+ + \Upsilon(X_n(s) \wedge Z_n)]\,\mathrm{d}s
\end{aligned}
$$

for each $t \ge 0$, where $X_n(0), Z_n \in \mathbb{R}^K$ are random vectors, $Y_{n,i} \in D^K$ for each $i = 1, \ldots, K$ are stochastic processes, and $P_n \in \mathcal{M}^K$. If the sequences $\{X_n(0) : n \ge 1\}$, $\{Z_n : n \ge 1\}$, and $\{Y_{n,i} : n \ge 1\}$ are stochastically bounded in $\mathbb{R}^K$ and D^K, respectively, for $i = 1, \ldots, K$, then the sequence $\{X_n : n \ge 1\}$ is stochastically bounded in D^K.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

Proof. We first note that

$$\int_0^t [(I - P_n(X_n(s))^\top)\Theta(X_n(s) - Z_n)^+ + \Upsilon(X_n(s) \wedge Z_n)]\,\mathrm{d}s$$
$$= \int_0^t [(I - P_n(X_n(s))^\top)\Theta(X_n(s) - Z_n)^+ - \Upsilon(X_n(s) - Z_n)^-]\,\mathrm{d}s + \Upsilon Z_n t.$$

Then, it follows that

$$\|X_n(t)\| \le \|X_n(0)\| + \|Y_{n,1}\|_T + \cdots + \|Y_{n,K}\|_T + (\|\Upsilon\| + (K^2\|\Theta\|) \vee \|\Upsilon\|)\|Z_n\|T$$
$$+ ((K^2\|\Theta\|) \vee \|\Upsilon\|) \int_0^t \|X_n(s)\|\,\mathrm{d}s.$$

The claim follows by applying Gronwall's inequality.

Lemma 4. (SB of $\{\bar{\boldsymbol{X}}^\lambda\}$ in the QED regime.) *Under the assumptions of Theorem 1, the sequences of K-dimensional $\boldsymbol{F}$-martingale vectors $\{\hat{\boldsymbol{M}}_A^\lambda\}$, $\{\hat{\boldsymbol{M}}_{L,0}^\lambda\}$, and $\{\hat{\boldsymbol{M}}_S^\lambda\}$, and the sequence of $(K \times K)$-dimensional martingale matrices $\{\hat{\boldsymbol{M}}_L^\lambda\}$, are all stochastically bounded, and, thus, the sequence of processes $\{\bar{\boldsymbol{X}}^\lambda\}$ is stochastically bounded.*

Proof. By Lemma 5.3 of Pang *et al*. (2007), SB for a sequence of random vectors in D^K is equivalent to SB for each sequence of individual components in D, and this also easily generalizes to SB of a sequence of random matrices in $D([0,\infty), \mathbb{R}^{K\times K})$. Thus, it suffices to prove the SB of the sequences of martingales $\{\hat{M}_{A,i}^\lambda\}$, $\{\hat{M}_{L,i,j}^\lambda\}$, and $\{\hat{M}_{S,i}^\lambda\}$ for each i, j. Moreover, by the criteria of SB of square-integrable martingales given in Lemma 5.9 of Pang *et al*. (2007), we only need to show that the corresponding sequences of random variables $\{\langle\hat{M}_{A,i}^\lambda\rangle(T)\}$, $\{\langle\hat{M}_{L,i,j}^\lambda\rangle(T)\}$, and $\{\langle\hat{M}_{S,i}^\lambda\rangle(T)\}$ are stochastic bounded in $\mathbb{R}$ for each i, j and $T > 0$.

First, $\langle\hat{M}_{A,i}^\lambda\rangle(T) = \lambda^{-1}\int_0^T \lambda_i(t)\,\mathrm{d}t$, so the SB of $\{\langle\hat{M}_{A,i}^\lambda\rangle(T)\}$ follows from (4). Second,

$$\langle\hat{M}_{L,i,j}^\lambda\rangle(T) = \theta_i \int_0^T p_{ij}^\lambda(r\bar{\boldsymbol{X}}^\lambda(s-))(\bar{X}_i^\lambda(s) - \bar{N}_j^\lambda)^+\,\mathrm{d}s$$
$$\le \theta_i(\bar{X}_i^\lambda(0) + \bar{N}_i^\lambda)T + \sum_{i=1}^K \frac{1}{\lambda} A_i\left(\int_0^T \lambda_i(s)\,\mathrm{d}s\right).$$

Hence, by the assumptions in (2) and (4) and the initial conditions, $p_{ij}^\lambda(\cdot) \le 1$, and by the weak law of large numbers for Poisson processes we can conclude the SB of $\{\langle\hat{M}_{L,i,j}^\lambda\rangle(T)\}$. Third, $\langle\hat{M}_{S,i}^\lambda\rangle(T) = \mu_i\int_0^T(\bar{X}_i^\lambda(t) \wedge \bar{N}_i^\lambda)\,\mathrm{d}t \le \mu_i\bar{N}_i^\lambda T$, which implies the SB of $\{\langle\hat{M}_{S,i}^\lambda\rangle(T)\}$. Next, we will prove the SB of $\{\bar{\boldsymbol{X}}^\lambda\}$. By the matrix-form representation of $\bar{\boldsymbol{X}}^\lambda$ in (39), and applying Lemma 3, we only need to check that the sequence of martingales in D^K, $\{\bar{\boldsymbol{M}}_A^\lambda + \sum_{i=1}^K[\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\bar{\boldsymbol{M}}_L - \bar{\boldsymbol{M}}_L\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1} - \bar{\boldsymbol{M}}_{L,0}^\lambda - \bar{\boldsymbol{M}}_S^\lambda\}$, is stochastically bounded. Indeed, this follows from the preservation of SB in vector and matrix forms and in sums of random elements in D^K, and $\bar{\boldsymbol{M}}_A^\lambda = \lambda^{-1/2}\hat{\boldsymbol{M}}_A^\lambda$ implies that the sequence $\{\bar{\boldsymbol{M}}_A^\lambda\}$ is stochastically bounded, and similarly for the other sequences of fluid-scaled martingales.

Proof of Theorem 1. We will apply the CMT to the mapping defined in (40) together with the convergence of the fluid-scaled martingales

$$(\bar{\boldsymbol{M}}_A^\lambda, \bar{\boldsymbol{M}}_L^\lambda, \bar{\boldsymbol{M}}_{L,0}^\lambda, \bar{\boldsymbol{M}}_S^\lambda) \xrightarrow{\mathrm{D}} (\mathbf{0}, \boldsymbol{O}, \mathbf{0}, \mathbf{0}) \quad \text{in } (D^K \times D^{K\times K} \times D^K \times D^K, J_1) \text{ as } \lambda \to \infty, \tag{41}$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

where $\mathbf{0}$ is the K-dimensional zero vector and $\boldsymbol{O}$ is the $(K \times K)$-dimensional zero matrix. The convergence in (41) follows from the FWLLN from SB, Lemma 5.10 of Pang *et al*. (2007). That result covers the case of vectors and can be easily generalized to the case of matrices.

A.3. Proof of the diffusion limits in the QED regime

Proof of Theorem 2. First, we have the following martingale representation of the diffusion-scaled queue-length processes $\hat{\boldsymbol{X}}^\lambda$ from (38):

$$\hat{X}^\lambda(t) = \hat{X}^\lambda(0) + \hat{\Lambda}^\lambda(t) - \Upsilon\hat{N}^\lambda t - \int_0^t [(I - \boldsymbol{P}^\lambda(\lambda\bar{\boldsymbol{X}}^\lambda(s-))^\top)\Theta(\hat{\boldsymbol{X}}^\lambda(s))^+ + \Upsilon(\hat{\boldsymbol{X}}^\lambda)^-]\,\mathrm{d}s$$
$$+ \hat{\boldsymbol{M}}_A^\lambda(t) - \hat{\boldsymbol{M}}_{L,0}^\lambda(t) - \hat{\boldsymbol{M}}_S^\lambda(t) + \sum_{i=1}^K [\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\hat{\boldsymbol{M}}_L^\lambda(t) - \hat{\boldsymbol{M}}_L^\lambda(t)\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1}. \tag{42}$$

Here $\hat{\Lambda}^\lambda(t) = (\sqrt{\lambda}(\lambda^{-1}\int_0^t \lambda_i(s)\,\mathrm{d}s - a_i t))^\top_{i=1,\dots,K}$, $\hat{\boldsymbol{N}}^\lambda = (\hat{N}_1^\lambda, \dots, \hat{N}_K^\lambda)^\top$, the martingale vector $\hat{\boldsymbol{M}}_A^\lambda(t) = (\hat{M}_{A,1}^\lambda(t), \dots, \hat{M}_{A,K}^\lambda(t))^\top$, and similarly for the martingale vectors $\hat{\boldsymbol{M}}_{L,0}^\lambda(t)$, $\hat{\boldsymbol{M}}_S^\lambda(t)$ and the martingale matrix $\hat{\boldsymbol{M}}_L^\lambda(t)$.

We again follow the martingale proof procedure to prove the diffusion limit by applying the CMT. The mapping defined from the prelimit diffusion-scaled processes in (42) is different from that in Lemma 2, and is given by the mapping $\psi\colon D^K \times D^K \times \mathbb{R}^K \times \mathbb{R}^K \times \mathcal{M}^K \to D^K$ that maps (y, z, a, b, P) into $x := \psi(y, z, a, b, P)$, i.e.

$$x(t) = b + y(t) - \int_0^t [(I - P(z(s))^\top)\Theta(x(s) - a)^+ + \Upsilon(x(s) \wedge a)]\,\mathrm{d}s, \qquad t \geq 0,$$

where $\mathcal{M}^K \ni P(\cdot)\colon \mathbb{R}^K \to [0,1]^{K\times K}$ is Lipschitz. A similar argument as Lemma 2 shows that this mapping is continuous in the space (D^K, J_1) and if $y, z \in C^K$ then $x \in C^K$. Next, we show the convergence of diffusion-scaled martingales

$$(\hat{\boldsymbol{M}}_A^\lambda, \hat{\boldsymbol{M}}_L^\lambda, \hat{\boldsymbol{M}}_{L,0}^\lambda, \hat{\boldsymbol{M}}_S^\lambda) \xrightarrow{\mathrm{D}} (\boldsymbol{B}_a, \boldsymbol{O}, \mathbf{0}, \boldsymbol{B}_s) \quad \text{in } (D^K \times D^{K\times K} \times D^K \times D^K, J_1) \tag{43}$$

as $\lambda \to \infty$, where $\boldsymbol{B}_a$ and $\boldsymbol{B}_s$ are both K-dimensional Brownian motions with mean $\mathbf{0}$, and covariance matrices $\Gamma_a = \mathrm{diag}\{a_1, \dots, a_K\}$ and $\Gamma_s = \mathrm{diag}\{\mu_1\nu_1, \dots, \mu_K\nu_K\}$, respectively. This follows from the fact that the square-integrable martingales $\hat{M}_{A,i}^\lambda(t)$, $\hat{M}_{L,i,j}^\lambda(t)$, $\hat{M}_{L,i,0}^\lambda(t)$, and $\hat{M}_{S,i}^\lambda(t)$ have quadratic variations

$$\langle\hat{M}_{A,i}^\lambda\rangle(t) = \frac{1}{\lambda}\int_0^t \lambda_i(s)\,\mathrm{d}s \to a_i t, \qquad \langle\hat{M}_{S,i}^\lambda\rangle(t) = \mu_i\int_0^t (\bar{X}_i^\lambda(s) \wedge \bar{N}_i^\lambda)\,\mathrm{d}s \xrightarrow{\mathrm{D}} \mu_i\nu_i t, \tag{44}$$

$$\langle\hat{M}_{L,i,j}^\lambda\rangle(t) = \theta_i\int_0^t p_{ij}^\lambda(\lambda\bar{\boldsymbol{X}}^\lambda(s-))(\bar{X}_i^\lambda(s) - \bar{N}_i^\lambda)^+\,\mathrm{d}s \xrightarrow{\mathrm{D}} 0, \tag{45}$$

as $\lambda \to \infty$ for each $t \geq 0$ and $i = 1, \dots, K$, $j = 0, 1, \dots, K$. Equations (44) and (45) follow from the heavy-traffic QED assumption (ii) and the SB of $\{\hat{\boldsymbol{X}}^\lambda\}$ because, by the FWLLN for a sequence of stochastic bounded processes, $\bar{\boldsymbol{X}}^\lambda = \lambda^{-1/2}\hat{\boldsymbol{X}}^\lambda + \lambda^{-1}\boldsymbol{N}^\lambda \xrightarrow{\mathrm{D}} \mathbf{0} + \boldsymbol{\nu} = \boldsymbol{\nu}$ as $\lambda \to \infty$. Finally, (43) follows from the FCLT for multidimensional square-integrable martingales, Theorem 8.1 of Pang *et al*. (2007).

Proof of Corollary 1. To prove the claim that $\hat{X}^{(1)}(t) \geq \hat{X}^{(2)}(t)$ for all $t \geq 0$, it suffices to argue that, given $\hat{X}^{(1)}(t) \geq \hat{X}^{(2)}(t)$ for some $t \geq 0$, the same inequality will hold at $t + \mathrm{d}t$ for

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

some (small) $\mathrm{d}t > 0$. Consider the ith component. From (14) we have

$$\begin{aligned}
&\mathrm{d}\hat{X}_i^{(1)}(t) - \mathrm{d}\hat{X}_i^{(2)}(t) \\
&\quad = \sum_{k=1}^{K} \theta_k (\bar{p}_{ki}^{(1)}(\boldsymbol{\nu})(\hat{X}_k^{(1)}(s))^+ - \bar{p}_{ki}^{(2)}(\boldsymbol{\nu})(\hat{X}_k^{(2)}(s))^+)\,\mathrm{d}t \\
&\qquad - \theta_i((\hat{X}_i^{(1)}(t))^+ - (\hat{X}_i^{(2)}(t))^+)\,\mathrm{d}t + \mu_i((\hat{X}_i^{(1)}(t))^- - (\hat{X}_i^2(t))^-)\,\mathrm{d}t. \qquad (46)
\end{aligned}$$

There are two cases. First, if $\hat{X}_i^{(1)}(t) > \hat{X}_i^{(2)}(t)$ then we are guaranteed to have $\hat{X}_i^{(1)}(t+\mathrm{d}t) \geq \hat{X}_i^{(2)}(t+\mathrm{d}t)$ for sufficiently small $\mathrm{d}t$, since $\hat{X}_i^{(1)}(t)$ and $\hat{X}_i^{(2)}(t)$ are both continuous in t. Second, if $\hat{X}_i^{(1)}(t) = \hat{X}_i^{(2)}(t)$ then (46) reduces to

$$\mathrm{d}\hat{X}_i^{(1)}(t) - \mathrm{d}\hat{X}_i^{(2)}(t) = \sum_{k=1}^{K} \theta_k (\bar{p}_{ki}^{(1)}(\boldsymbol{\nu}) - \bar{p}_{ki}^{(2)}(\boldsymbol{\nu}))(\hat{X}_k^{(1)}(s))^+ \mathrm{d}t \geq 0; \qquad (47)$$

hence, $\hat{X}_i^{(1)}(t+\mathrm{d}t) \geq \hat{Y}_i^{(2)}(t+\mathrm{d}t)$. In fact, applying the same argument, we can strengthen it to (16). For instance, suppose that $\hat{X}_i^{(1)}(t) = \hat{X}_i^{(2)}(t) < 0$ at t, i.e. $(\hat{X}_i^{(1)}(t))^- = (\hat{X}_i^{(2)}(t))^-$. Then (46) again reduces to (47). Hence,

$$(\hat{X}_i^{(1)}(t))^- - \mathrm{d}\hat{X}_i^{(1)}(t) \leq (\hat{X}_i^{(2)}(t))^- - \mathrm{d}\hat{X}_i^{(2)}(t),$$

which simplifies to $-\hat{X}_i^{(1)}(t+\mathrm{d}t) \leq -\hat{X}_i^{(2)}(t+\mathrm{d}t)$ or $(\hat{X}_i^{(1)}(t+\mathrm{d}t))^- \leq (\hat{X}_i^{(2)}(t+\mathrm{d}t))^-$, taking into account the facts that $\hat{X}_i^{(1)}(t+\mathrm{d}t) < 0$ and $\hat{X}_i^{(2)}(t+\mathrm{d}t) < 0$, which follow from path continuity, along with a sufficiently small $\mathrm{d}t$.

The claim of $\hat{V}^{(1)}(t) \geq \hat{V}^{(2)}(t)$ follows from (16) and (15).

Proof of Theorem 3. First, we define the processes $A_i^\lambda(t)$, $L_{i,j}^\lambda(t)$, and $S_{i,j}^\lambda(t)$ as

$$A_i^\lambda(t) := A_i\left(\int_0^t \lambda_i(s)\,\mathrm{d}s\right), \qquad S_{i,j}^\lambda(t) := S_i\left(\mu_i \int_0^t (X_i(s) \wedge N_i)\,\mathrm{d}s\right),$$
$$L_{i,j}^\lambda(t) := L_{i,j}\left(\theta_i \int_0^t p_{ij}(\boldsymbol{X}(s-))(X_i(s) - N_i)^+\,\mathrm{d}s\right),$$

for each $i = 1, \ldots, K$, $j = 0, 1, \ldots, K$, and $t \geq 0$. Then, we define the following first passage time processes $\boldsymbol{V}_l^\lambda := (V_{l,1}^\lambda, \ldots, V_{l,K}^\lambda)$ and $\boldsymbol{V}_u^\lambda := (V_{u,1}^\lambda, \ldots, V_{u,K}^\lambda)$ as

$$\begin{aligned}
V_{l,i}^\lambda(t) &:= \inf\Bigg\{ s \geq 0 \;\Bigg|\; \sum_{j=1,\, j\neq i}^{K} L_{i,j}^\lambda(t+s) + L_{i,0}^\lambda(t+s) + S_{i,j}^\lambda(t+s) \\
&\qquad \geq X_i^\lambda(0) + A_i^\lambda(t) + \sum_{k=1,\, k\neq i}^{K} L_{k,i}^\lambda(t) - (N_i^\lambda - 1)\Bigg\} \\
&= \inf\Bigg\{ s \geq 0 \;\Bigg|\; \sum_{j=1,\, j\neq i}^{K} L_{i,j}^\lambda(t+s) + L_{i,0}^\lambda(t+s) + S_{i,j}^\lambda(t+s) \\
&\qquad \geq X_i^\lambda(t) + \sum_{j=1,\, j\neq i}^{K} L_{i,j}^\lambda(t) + L_{i,0}^\lambda(t) + S_{i,j}^\lambda(t) - (N_i^\lambda - 1)\Bigg\},
\end{aligned} \qquad (48)$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

$$V_{u,i}^{\lambda}(t) := \inf\Bigg\{ s \ge 0 \;\Bigg|\; \sum_{j=1,\, j\neq i}^{K} L_{i,j}^{\lambda}(t) + L_{i,0}^{\lambda}(t) + S_{i,j}^{\lambda}(t+s) \ge X_i^{\lambda}(0) + A_i^{\lambda}(t) + \sum_{k=1,\, k\neq i}^{K} L_{k,i}^{\lambda}(t) - (N_i^{\lambda}-1) \Bigg\}$$
$$= \inf\{ s \ge t \mid S_{i,j}^{\lambda}(t+s) \ge X_i^{\lambda}(t) + S_{i,j}^{\lambda}(t) - (N_i^{\lambda}-1) \}, \tag{49}$$

where the second equalities in (48) and (49) follow from identity (1). The first passage time process $V_{l,i}^{\lambda}(t)$ is the first time after time t when the server pool i frees up a server if the arrival process of class-i jobs is stopped at time t and the switchovers from the other classes to class i are also stopped at time t, while the switchover from class i to the other classes or leaving the system without receiving service are allowed after time t. This clearly provides a lower bound for $V_i^{\lambda}(t)$. Since the second line of (48), $X_i^{\lambda}(0) + A_i^{\lambda}(t) + \sum_{k=1,\, k\neq i}^{K} L_{k,i}^{\lambda}(t) - (N_i^{\lambda}-1)$, is nondecreasing in t, the first passage time process $V_{l,i}^{\lambda}(t)$ has sample paths in D. The first passage time process $V_{u,i}^{\lambda}(t)$ is the first time after time t when the server pool i frees up a server if the arrival process of class i jobs, and switchovers either from class i to the other classes or from the other classes to class i or leaving the system without receiving service, are all stopped after time t. It is also clear as an upper bound for $V_i^{\lambda}(t)$. However, we note that the term on the right-hand side of (49), $X_i^{\lambda}(0) + A_i^{\lambda}(t) + \sum_{k=1,\, k\neq i}^{K} L_{k,i}^{\lambda}(t) - \sum_{j=1,\, j\neq i}^{K} L_{i,j}^{\lambda}(t) - L_{i,0}^{\lambda}(t) - (N_i^{\lambda}-1)$, is not necessarily nondecreasing in t and, thus, the first passage time process $V_{u,i}^{\lambda}(t)$ does not necessarily have sample paths in D. We can use the trick of linear interpolation as in the proof of Theorem 3.1 of Talreja and Whitt (2009) to construct another process $\tilde{V}_{u,i}^{\lambda}(t)$ such that $V_{u,i}^{\lambda}(t) \le \tilde{V}_{u,i}^{\lambda}(t)$ for each $t \ge 0$, $\sup_{0\le t\le T} \sqrt{\lambda} |V_{u,i}^{\lambda}(t) - \tilde{V}_{u,i}^{\lambda}(t)| \to 0$ with probability 1 as $\lambda \to \infty$ for each $T > 0$, and $\tilde{V}_{u,i}^{\lambda}(t)$ has sample paths in D. The details are omitted.

In order to apply the corollary in Puhalskii (1994, p. 951) to prove (15), we define the following first passage time processes $\mathbf{Z}_l^{\lambda} = (Z_{l,1}^{\lambda}, \ldots, Z_{l,K}^{\lambda})^{\top}$ and $\mathbf{Z}_u^{\lambda} = (Z_{u,1}^{\lambda}, \ldots, Z_{u,K}^{\lambda})^{\top}$:

$$Z_{l,i}^{\lambda}(t) := \inf\Bigg\{ s \ge 0 \;\Bigg|\; \sum_{j=1,\, j\neq i}^{K} L_{i,j}^{\lambda}(s) + L_{i,0}^{\lambda}(s) + S_{i,j}^{\lambda}(s) \ge X_i^{\lambda}(t) + \sum_{j=1,\, j\neq i}^{K} L_{i,j}^{\lambda}(t) + L_{i,0}^{\lambda}(t) + S_{i,j}^{\lambda}(t) - (N_i^{\lambda}-1) \Bigg\},$$
$$Z_{u,i}^{\lambda}(t) = \inf\{ s \ge 0 \mid S_{i,j}^{\lambda}(s) \ge X_i^{\lambda}(t) + S_{i,j}^{\lambda}(t) - (N_i^{\lambda}-1) \}.$$

It is clear that $V_{l,i}^{\lambda}(t) = (Z_{l,i}^{\lambda}(t) - t)^+$ and $V_{u,i}^{\lambda}(t) = (Z_{u,i}^{\lambda}(t) - t)^+$ for each $t \ge 0$.

By (43) and the CMT, we have

$$\sqrt{\lambda}\Bigg(\sum_{j=1,\, j\neq i}^{K} \frac{1}{\lambda} L_{i,j}^{\lambda}(s) + \frac{1}{\lambda} L_{i,0}^{\lambda}(s) + \frac{1}{\lambda} S_{i,j}^{\lambda}(s) - \nu_i \mu_i s \Bigg) \xrightarrow{\mathrm{D}} \sqrt{\mu_i \nu_i}\, B_{s,i}(s),$$
$$\sqrt{\lambda}\Bigg(\frac{1}{\lambda} S_{i,j}^{\lambda}(s) - \nu_i \mu_i s \Bigg) \xrightarrow{\mathrm{D}} \sqrt{\mu_i \nu_i}\, B_{s,i}(s),$$

and, together with (13) in Theorem 2,

$$\sqrt{\lambda}\Bigg(\frac{1}{\lambda} X_i^{\lambda}(t) + \sum_{j=1,\, j\neq i}^{K} \frac{1}{\lambda} L_{i,j}^{\lambda}(t) + \frac{1}{\lambda} L_{i,0}^{\lambda}(t) + \frac{1}{\lambda} S_{i,j}^{\lambda}(t) - \frac{1}{\lambda}(N_i^{\lambda}-1) - \nu_i \mu_i t \Bigg)$$
$$\xrightarrow{\mathrm{D}} \hat{X}_i(t) + \sqrt{\mu_i \nu_i}\, B_{s,i}(t)$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

and

$$\sqrt{\lambda}\left(\frac{1}{\lambda}X_i^\lambda(t)+\frac{1}{\lambda}S_{i,j}^\lambda(t)-\frac{1}{\lambda}(N_i^\lambda-1)-\nu_i\mu_i t\right)\xrightarrow{\mathrm{D}}\hat{X}_i(t)+\sqrt{\mu_i\nu_i}B_{s,i}(t),$$

where the convergence is the space (D, J_1) as $\lambda\to\infty$.

Then, by the corollary in Puhalskii (1994, p. 951), we obtain

$$\left(\sqrt{\lambda}\left(\frac{1}{\lambda}Z_{l,i}^\lambda(t)-t\right),\sqrt{\lambda}\left(\frac{1}{\lambda}Z_{l,i}^\lambda(t)-t\right)\right)\xrightarrow{\mathrm{D}}(\hat{Z}_{l,i}(t),\hat{Z}_{u,i}(t))\quad\text{in } D^2$$

as $\lambda\to\infty$, where $\hat{Z}_{l,i}(t)=\hat{Z}_{u,i}(t)=\hat{X}_i(t)/(\mu_i\nu_i)$, $t\ge 0$. Finally, the theorem is proved by applying the CMT to the mapping $x\to x^+\colon\mathbb{R}\to\mathbb{R}_+$.

A.4. Proof of the fluid limit in the ED regime

The prelimit fluid-scaled queue-length processes $\bar{\boldsymbol{X}}^\lambda$ have the same representation as (39) in the QED regime. The proof is very similar to the proof of Theorem 1 by applying the CMT to the mapping defined in Lemma 2 together with the convergence of fluid-scaled martingales. We omit the details.

A.5. Proof of the diffusion limit in the ED regime

Proof of Theorem 5. First, by Lemma 4, there exists $t_0>0$ such that $\boldsymbol{x}(t)>\boldsymbol{\nu}$ for each $t>t_0$, and, under the heavy-traffic ED assumption, there exists λ_0 such that, for any $\lambda>\lambda_0$, $\inf_{0<s\le T}X_i^\lambda(s)>N_i^\lambda$, $i=1,\dots,K$. From (19) and (39), we have the following martingale representation for the diffusion-scaled queue-length processes $\hat{\boldsymbol{X}}^\lambda$:

$$\begin{aligned}\hat{\boldsymbol{X}}^\lambda(t)&=\hat{\boldsymbol{X}}^\lambda(0)+\hat{\boldsymbol{M}}_A^\lambda(t)+\sum_{i=1}^K[\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\hat{\boldsymbol{M}}_L^\lambda(t)-\hat{\boldsymbol{M}}_L^\lambda(t)\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1}-\hat{\boldsymbol{M}}_{L,0}^\lambda(t)-\hat{\boldsymbol{M}}_S^\lambda(t)\\&\quad-\int_0^t[(I-\boldsymbol{P}^\lambda(\lambda\bar{\boldsymbol{X}}^\lambda(s-))^\top)\Theta\hat{\boldsymbol{X}}^\lambda(s)-\hat{\boldsymbol{P}}^\lambda(\bar{\boldsymbol{X}}^\lambda(s-))\Theta(\boldsymbol{x}(s)-\boldsymbol{\nu})]\,\mathrm{d}s\\&\quad-\int_0^t\sqrt{\lambda}[(I-\bar{\boldsymbol{P}}(\bar{\boldsymbol{X}}^\lambda(s-))^\top)-(I-\bar{\boldsymbol{P}}(\boldsymbol{x}(s))^\top)]\Theta(\boldsymbol{x}(s)-\boldsymbol{\nu})\,\mathrm{d}s\end{aligned}\tag{50}$$

for $t>t_0$ and $\lambda>\lambda_0$. Here $\hat{\boldsymbol{P}}^\lambda(\cdot)=[\hat{p}_{ij}^\lambda(\cdot)]_{i,j=1}^K\in\mathcal{M}^K$.

Second, we define another sequence of processes $\hat{\boldsymbol{Y}}^\lambda$ with $\hat{\boldsymbol{Y}}^\lambda(0)=\hat{\boldsymbol{X}}^\lambda(0)$, and

$$\begin{aligned}\hat{\boldsymbol{Y}}^\lambda(t)&=\hat{\boldsymbol{X}}^\lambda(0)+\hat{\boldsymbol{M}}_A^\lambda(t)-\hat{\boldsymbol{M}}_{L,0}^\lambda(t)-\hat{\boldsymbol{M}}_S^\lambda(t)+\sum_{i=1}^K[\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\hat{\boldsymbol{M}}_L^\lambda(t)-\hat{\boldsymbol{M}}_L^\lambda(t)\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1}\\&\quad-\int_0^t[(I-\boldsymbol{P}^\lambda(\lambda\bar{\boldsymbol{X}}^\lambda(s-))^\top)\Theta\hat{\boldsymbol{Y}}^\lambda(s)-\hat{\boldsymbol{P}}^\lambda(\bar{\boldsymbol{X}}^\lambda(s-))\Theta(\boldsymbol{x}(s)-\boldsymbol{\nu})]\,\mathrm{d}s.\end{aligned}\tag{51}$$

We then show that the difference between $\hat{\boldsymbol{X}}^\lambda$ and $\hat{\boldsymbol{Y}}^\lambda$ is asymptotically negligible as $\lambda\to\infty$. By (50) and (51), we have

$$\begin{aligned}\hat{\boldsymbol{X}}^\lambda(t)-\hat{\boldsymbol{Y}}^\lambda(t)&=\int_0^t(I-\boldsymbol{P}^\lambda(\lambda\bar{\boldsymbol{X}}^\lambda(s-))^\top)\Theta(\hat{\boldsymbol{X}}^\lambda(s)-\hat{\boldsymbol{Y}}^\lambda(s))\,\mathrm{d}s\\&\quad+\int_0^t\sqrt{\lambda}[(I-\bar{\boldsymbol{P}}(\bar{\boldsymbol{X}}^\lambda(s-))^\top)-(I-\bar{\boldsymbol{P}}(\boldsymbol{x}(s))^\top)]\Theta(\boldsymbol{x}(s)-\boldsymbol{\nu})\,\mathrm{d}s.\end{aligned}$$

By the Lipschitz property of $\bar{\boldsymbol{P}}$ and the convergence $\bar{\boldsymbol{X}}^\lambda\xrightarrow{\mathrm{D}}\boldsymbol{x}$ as $\lambda\to\infty$, we have $\|\hat{\boldsymbol{X}}^\lambda(t)-\hat{\boldsymbol{Y}}^\lambda(t)\|\xrightarrow{\mathrm{D}}0$ as $\lambda\to\infty$.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

It therefore remains to show the convergence of the processes $\hat{\boldsymbol{Y}}^\lambda$. The corresponding K-dimensional integral mapping $\varphi\colon D^K \times D^K \times C^K \times \mathbb{R}^K \times \mathcal{M}^K \times \mathcal{M}^K \to D^K$ that maps (y, z, w, b, P, Q) into $x := \varphi(y, z, w, b, P, Q)$ is defined by

$$x(t) = b + y(t) - \int_0^t [(I - P(z(s))^\top)\Theta x(s) - Q(z(s))\Theta(w(s) - \nu)]\,\mathrm{d}s, \qquad t \geq 0,$$

where $\mathcal{M}^K \ni P(\cdot), Q(\cdot)\colon \mathbb{R}^K \to [0,1]^{K\times K}$ are Lipschitz. A similar argument as used in the proof of Lemma 2 shows that this mapping is continuous in the space (D^K, J_1) and if $y, z \in C^K$ then $x \in C^K$. The sequences of diffusion-scaled $\boldsymbol{F}$-martingales converge,

$$(\hat{\boldsymbol{M}}_A^\lambda, \hat{\boldsymbol{M}}_L^\lambda, \hat{\boldsymbol{M}}_{L,0}^\lambda, \hat{\boldsymbol{M}}_S^\lambda) \xrightarrow{\mathrm{D}} (\hat{\boldsymbol{B}}_A, \hat{\boldsymbol{B}}_L, \hat{\boldsymbol{B}}_{L,0}, \hat{\boldsymbol{B}}_S),$$

in $(D^K \times D^{K\times K} \times D^K \times D^K, J_1)$ as $\lambda \to \infty$, where

$$\begin{aligned}
\hat{\boldsymbol{B}}_A(t) &:= \left(\hat{B}_{A,i}\left(\int_0^t a_i(s)\,\mathrm{d}s\right)\right)^\top_{i=1,\dots,K},\\
\hat{\boldsymbol{B}}_S(t) &:= (\hat{B}_{S,i}(\mu_i\nu_i t))^\top_{i=1,\dots,K},\\
\hat{\boldsymbol{B}}_L(t) &:= \left(\hat{B}_{L,i,j}\left(\theta_i\int_0^t \bar{p}_{ij}(\boldsymbol{x}(s))(x_i(s) - \nu_i)\,\mathrm{d}s\right)\right)_{i,j=1,\dots,K},\\
\hat{\boldsymbol{B}}_{L,0}(t) &:= \left(\hat{B}_{L,i,0}\left(\theta_i\int_0^t \bar{p}_{i0}(\boldsymbol{x}(s))(x_i(s) - \nu_i)\,\mathrm{d}s\right)\right)^\top_{i=1,\dots,K},
\end{aligned}$$

$\hat{B}_{A,i}$, $\hat{B}_{L,i,j}$, and $\hat{B}_{S,i}$ are mutually independent standard Brownian motions, and $\boldsymbol{x}(t)$ is defined in (19). This follows by applying the FCLT for multidimensional square-integrable martingales because $\hat{M}_{A,i}^\lambda(t)$, $\hat{M}_{L,i,j}^\lambda(t)$, $\hat{M}_{L,i,0}^\lambda(t)$, and $\hat{M}_{S,i}^\lambda(t)$ have quadratic variations:

$$\begin{aligned}
\langle\hat{M}_{A,i}^\lambda\rangle(t) &= \frac{1}{\lambda}\int_0^t \lambda_i(s)\,\mathrm{d}s = \int_0^t a_i(s)\,\mathrm{d}s,\\
\langle\hat{M}_{S,i}^\lambda\rangle(t) &= \mu_i\int_0^t (\bar{X}_i^\lambda(s) \wedge \bar{N}_i^\lambda)\,\mathrm{d}s \xrightarrow{\mathrm{D}} \mu_i\nu_i t,
\end{aligned} \tag{52}$$

$$\begin{aligned}
\langle\hat{M}_{L,i,j}^\lambda\rangle(t) &= \theta_i\int_0^t p_{ij}^\lambda(\lambda\bar{\boldsymbol{X}}^\lambda(s-))(\bar{X}_i^\lambda(s) - \bar{N}_i^\lambda)^+\,\mathrm{d}s\\
&\xrightarrow{\mathrm{D}} \theta_i\int_0^t \bar{p}_{ij}(\boldsymbol{x}(s))(x_i(s) - \nu_i)\,\mathrm{d}s,
\end{aligned} \tag{53}$$

as $\lambda \to \infty$, for each $t \geq 0$ and $i = 1, \dots, K$, $j = 0, 1, \dots, K$. Equations (52) and (53) follow from the heavy-traffic ED assumption (17) and the SB of $\{\hat{\boldsymbol{X}}^\lambda\}$.

Finally, by applying the CMT to the mapping φ and the convergence of diffusion-scaled martingales, we obtain $\hat{\boldsymbol{Y}}^\lambda \xrightarrow{\mathrm{D}} \hat{\boldsymbol{Y}}$ in (D^K, J_1) as $\lambda \to \infty$, where

$$\begin{aligned}
\mathrm{d}\hat{\boldsymbol{Y}}(t) &= \hat{\boldsymbol{P}}(\boldsymbol{x}(s))\Theta(\boldsymbol{x}(t) - \boldsymbol{\nu})\,\mathrm{d}t - (I - \bar{\boldsymbol{P}}(\boldsymbol{x}(t))^\top)\Theta\hat{\boldsymbol{Y}}(t)\,\mathrm{d}t\\
&\quad + \mathrm{d}\left(\hat{\boldsymbol{B}}_A(t) + \sum_{i=1}^K [\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\hat{\boldsymbol{B}}_L(t) - \hat{\boldsymbol{B}}_L(t)\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1} - \hat{\boldsymbol{B}}_{L,0}(t) - \hat{\boldsymbol{B}}_S(t)\right),
\end{aligned}$$

and $\boldsymbol{x}(t)$ is defined in (19). It is easy to check that the Brownian terms here are equivalent in distribution to the Brownian motion $\boldsymbol{W}$ in (25).

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

Proof of Corollary 4. Here we give a proof by defining the diffusion-scaled process $\hat{\boldsymbol{X}}^\lambda$ in terms of centering around the steady state of the fluid limit process.

Define the diffusion-scaled processes $\hat{\boldsymbol{X}}^\lambda$ by

$$\hat{X}_i^\lambda(t) := \lambda^{-1/2}(X_i^\lambda(t) - (N_i^\lambda + \lambda q_i^*)), \qquad i = 1, \dots, K,$$

with q_i^* given in (23). Under the heavy-traffic ED assumption, for any given $\varepsilon \in (0, \max_i q_i^*)$ and $T > 0$, there exists λ_0 such that, for any $\lambda \geq \lambda_0$ and each i,

$$\inf_{0 \leq s \leq T} X_i^\lambda(s) \geq \lambda(v_i + q_i^* - \varepsilon) > N_i^\lambda.$$

Thus, for any $\lambda \geq \lambda_0$ and each i, we can write $\hat{\boldsymbol{X}}^\lambda$ in the following matrix form for all $\lambda > \lambda_0$:

$$\hat{\boldsymbol{X}}^\lambda(t) = \hat{\boldsymbol{X}}^\lambda(0) + \hat{\boldsymbol{M}}_A^\lambda(t) + \sum_{i=1}^K [\boldsymbol{E}^{(ii)}(\tilde{\boldsymbol{E}}\hat{\boldsymbol{M}}_L^\lambda(t) - \hat{\boldsymbol{M}}_L^\lambda(t)\tilde{\boldsymbol{E}})\boldsymbol{E}^{(ii)}]\mathbf{1}$$
$$- \hat{\boldsymbol{M}}_{L,0}^\lambda(t) - \hat{\boldsymbol{M}}_S^\lambda(t) - \int_0^t (I - \boldsymbol{P}^\top)\Theta\hat{\boldsymbol{X}}^\lambda(s)\,\mathrm{d}s. \tag{54}$$

The martingale terms in (54) converge as in the proof of Theorem 5, but

$$\langle \hat{M}_{L,i,j}^\lambda \rangle(t) = \theta_i p_{ij} \int_0^t (\bar{X}_i^\lambda(s) - \bar{N}_i^\lambda)^+ \,\mathrm{d}s \xrightarrow{\mathrm{D}} \theta_i p_{ij} q_i^* t, \qquad i = 1, \dots, K,\ j = 0, 1, \dots, K.$$

We can then apply the CMT to the simple mapping $(y, b, P) \to x\colon D^K \times \mathbb{R}^K \times \mathcal{M}^K \to D^K$, defined by the $x(t) = b + y(t) - \int_0^t (I - \boldsymbol{P}^\top)\Theta x(s)\,\mathrm{d}s$.

A.6. Proofs of the fluid and diffusion limits in the mixed regimes

Proof of Theorem 6. First, we have the same martingale representation of $\bar{\boldsymbol{X}}^\lambda$ as in (39). Second, we can show the SB of the multidimensional integral representation and the convergence of fluid-scaled martingales. Third, we can apply CMT to the mapping in Lemma 2 to conclude the convergence of $\bar{\boldsymbol{X}}^\lambda$. Last, the steady states are derived directly from the fluid limit.

Proof of Theorem 7. Here we only focus on the mixed QD and ED regime. First, as in (42) and (54), we have the following martingale representations of the processes $\hat{X}_i^\lambda$ for large λ and $i \in \tilde{\mathcal{K}}_1$, $\tilde{\mathcal{K}}_2$ and $\tilde{\mathcal{K}}_3$, separately:

$$\hat{X}_i^\lambda(t) = \hat{X}_i^\lambda(0) + \hat{M}_{A,i}^\lambda(t) + \sum_{k \in \tilde{\mathcal{K}}_2 \cup \tilde{\mathcal{K}}_3} \hat{M}_{L,k,i}^\lambda(t) - \hat{M}_{L,i,0}^\lambda(t) - \hat{M}_{S,i}^\lambda(t) - \mu_i \int_0^t \hat{X}_i^\lambda(s)\,\mathrm{d}s$$
$$+ \int_0^t \left(\sum_{k \in \tilde{\mathcal{K}}_2} \theta_k p_{ki}^\lambda (\hat{X}_k^\lambda(s))^+ + \sum_{k \in \tilde{\mathcal{K}}_3} \theta_k p_{ki}^\lambda \hat{X}_k^\lambda(s) \right) \mathrm{d}s \quad \text{for } i \in \tilde{\mathcal{K}}_1, \tag{55}$$

$$\hat{X}_i^\lambda(t) = \hat{X}_i^\lambda(0) + \hat{M}_{A,i}^\lambda(t) + \sum_{k \in \tilde{\mathcal{K}}_2 \cup \tilde{\mathcal{K}}_3,\, k \neq i} \hat{M}_{L,k,i}^\lambda(t) - \sum_{j=1,\, j \neq i}^K \hat{M}_{L,i,j}^\lambda(t) - \hat{M}_{L,i,0}^\lambda(t)$$
$$- \hat{M}_{S,i}^\lambda(t) - \mu_i \int_0^t (\hat{X}_i^\lambda(s) \wedge 0)\,\mathrm{d}s - \int_0^t \left(\theta_i (1 - p_{ii}^\lambda)(\hat{X}_i^\lambda(s))^+ \right) \mathrm{d}s$$
$$+ \int_0^t \left(\sum_{k \in \tilde{\mathcal{K}}_2,\, k \neq i} \theta_k p_{ki}^\lambda (\hat{X}_k^\lambda(s))^+ + \sum_{k \in \tilde{\mathcal{K}}_3} \theta_k p_{ki}^\lambda (\hat{X}_k^\lambda(s))^+ \right) \mathrm{d}s \quad \text{for } i \in \tilde{\mathcal{K}}_2, \tag{56}$$

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

$$\hat{X}_i^\lambda(t) = \hat{X}_i^\lambda(0) + \hat{M}_{A,i}^\lambda(t) + \sum_{k\in\tilde{\mathcal{K}}_2\cup\tilde{\mathcal{K}}_3,\, k\neq i} \hat{M}_{L,k,i}^\lambda(t) - \sum_{j=1,\, j\neq i}^{K} \hat{M}_{L,i,j}^\lambda(t) - \hat{M}_{L,i,0}^\lambda(t)$$
$$- \hat{M}_{S,i}^\lambda(t) + \int_0^t \left(\sum_{k\in\tilde{\mathcal{K}}_2} \theta_k p_{ki} (\hat{X}_k^\lambda(s))^+ + \sum_{k\in\tilde{\mathcal{K}}_3,\, k\neq i} \theta_k p_{ki} \hat{X}_k^\lambda(s) \right) \mathrm{d}s$$
$$- \theta_i(1 - p_{ii}) \int_0^t \hat{X}_i^\lambda(s)\, \mathrm{d}s \quad \text{for } i \in \tilde{\mathcal{K}}_3. \tag{57}$$

It is easy to see that the multidimensional integral mapping defined by (55)–(57) that maps the martingale terms and initial conditions to $\hat{\boldsymbol{X}}^\lambda$ is continuous in the Skorokhod J_1 topology, as in the proof of Lemma 2. It is also easy to check that $\hat{\boldsymbol{X}}^\lambda$ is SB. All the diffusion-scaled martingales converge, and only differ in the limits of their quadratic variations. In particular, for $i \in \tilde{\mathcal{K}}_3$, $\langle \hat{M}_{L,i,j}^\lambda \rangle(t) = \theta_i \int_0^t p_{ij}^\lambda (\bar{X}_i^\lambda(s) - \bar{N}_i^\lambda)^+ \,\mathrm{d}s \xrightarrow{\mathrm{D}} \theta_i \bar{p}_{ij}(x_i^* - \nu_i)t$, and $\langle \hat{M}_{L,i,j}^\lambda \rangle(t) \xrightarrow{\mathrm{D}} 0$ for $i \in \tilde{\mathcal{K}}_1 \cup \tilde{\mathcal{K}}_2$. The martingale limits contribute to the Brownian motion term in (35)–(37). Thus, the convergence follows by applying the CMT together with the convergence of the diffusion-scaled martingales. The convergence in the other mixed regimes follows similarly and is thus omitted.

Acknowledgements

The authors thank the anonymous referee for helpful comments on the paper. David Yao was supported in part by the NSF grant CMMI-0969328.

References

Berman, A. and Plemmons, R. J. (1979). *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd edn. John Wiley, New York.

Bonald, T. *et al*. (2009). Flow-level performance and capacity of wireless networks with user mobility. *Queueing Systems* **63,** 131–164.

Borst, S., Mandelbaum, A. and Reiman, M. I. (2004). Dimensioning large call centers. *Operat. Res.* **52,** 17–34.

Borst, S., Proutiere, A. and Hegde, N. (2006). Capacity of wireless data networks with intra- and inter-cell mobility. In *Proc. IEEE INFOCOM 2006* (Barcelona, Spain), 12pp.

Bramson, M., Lu, Y. and Prabhakar, B. (2010). Randomized load balancing with general service time distributions. *ACM SIGMETRICS Performance Evaluation Rev.* **38,** 275–286.

Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization* (Appl. Math. **46**). Springer, New York.

Dieker, A. B. and Gao. X. (2013). Positive recurrence of piecewise Ornstein-Uhlenbeck processes and common quadratic Lyapunov function. *Ann. Appl. Prob.* **23,** 1291–1317.

Erlang, A. K. (1948). On the rational determination of the number of circuits. In *The Life and Works of A. K. Erlang*, eds E. Brockmeyer, H. L. Halstrom and A. Jensen, The Copenhagen Telephone Company, Copenhagen, Denmark.

Fleming, P. J., Stolyar, A. and Simon, B. (1995). Heavy traffic limit for a mobile phone system loss model. In *Proc. Nashville ACM Telecommunications Conf.*

Ganesh, A. *et al*. (2010). Load balancing via random local research in closed and open systems. *ACM SIGMETRICS Performance Evaluation Rev*. **38,** 287–298.

Garnett, O., Mandelbaum, A. and Reiman, M. I. (2002). Designing a call center with impatient customers. *Manufacturing Service Operat. Manag.* **4,** 208–227.

Grossglauser, M. and Tse, D. N. C. (2002). Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Trans. Networking* **10,** 477–486.

Gurvich, I. and Whitt, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Math. Operat. Res.* **34,** 363–396.

Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operat. Res.* **29,** 567–588.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press

He, S. and Dai, J. (2013). Many-server queues with customer abandonment: numerical analysis of their diffusion models. To appear in *Stoch. Systems.*

Kang, W. and Pang, G. (2013). Fluid limit of a multiclass many-server queueing network with abandonment and feedback. Submitted.

Kang, W. and Ramanan, K. (2010). Fluid limits of many-server queues with reneging. *Ann. Appl. Prob.* **20,** 2204–2260.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd edn. Springer, New York.

Mandelbaum, A., Massey, W. A. and Reiman, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems* **30,** 149–201.

Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Prob. Surveys* **4,** 193–267.

Perry, O. and Whitt, W. (2013). A fluid limit for an overloaded X model via a stochastic averaging principle. *Math. Operat. Res.* **38,** 294–349.

Puhalskii, A. (1994). On the invariance principle for the first passage time. *Math. Operat. Res.* **19,** 946–954.

Saure, D., Glynn, P. W. and Zeevi, A. (2009). A linear programming algorithm for computing the stationary distribution of semimartingale reflected Brownian motion. Working paper.

Simatos, F. and Tibi, D. (2010). Spatial homogenization in a stochastic network with mobility. *Ann. Appl. Prob*. **20,** 312–355.

Talreja, R. and Whitt, W. (2009). Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Prob.* **19,** 2137–2175.

Whitt, W. (2002). *Stochastic-Process Limits.* Springer, New York.

Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Manag. Sci.* **50,** 1449–1461.

https://doi.org/10.1239/aap/1377868533 Published online by Cambridge University Press