OPERATIONS RESEARCH
http://dx.doi.org/10.1287/opre.1120.1070ec

© 2012 INFORMS

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Overflow Networks: Approximations and Implications to Call-Center Outsourcing" by Itai Gurvich and Ohad Perry, *Operations Research*, http://dx.doi.org/10.1287/opre.1120.1070.

Proofs and extensions

This e-companion is divided into two sections. In §EC.1 we prove all the results that appears in the body of the paper. In section EC.2 we consider the extension of the base model to a multi-class setting as the one in Figure 1(b) of the paper.

EC.1. Proofs

We start by introducing some additional notational conventions. For two random variables X and Y we write $X \leq_{st} Y$ when stochastic ordering holds in the standard sense. Namely, when $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$ for every non-negative non-decreasing function $f(\cdot)$ for which the expectations are defined. If X and Y are two $\mathcal{D}$-valued stochastic processes we will write $\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0)\}$ to denote the fact that the processes are ordered. In other words, that there exists a construction of the sample paths of X and Y such that, almost surely, $X(t) \leq Y(t)$ for all $t \geq 0$.

The rest of the section is divided into two subsections. Subsection EC.1.1 establishes important properties of the (sequence of) availability processes D_I^λ, as defined in (8). Building on these, we then proceed to §EC.1.2 where we prove the main results sated in §4. The proofs of several auxiliary lemmas are relegated to §EC.1.3.

EC.1.1. The Availability Process

We first establish upper and lower bounds (in appropriate probabilistic sense) for the process D_I^λ, by relating it to simple $M/M/1$ queues; see Lemmas EC.1.1 and EC.1.3. These bounds are then used to establish the pointwise stationarity of D_I^λ – see Theorem EC.1.4. A FCLT via an AP for an appropriately scaled version of the related cumulative process C_I^λ in (9) is stated and proved in Theorem EC.1.6.

We start with a useful comparison result. For the following, let $Q_\epsilon^+ = (Q_\epsilon^+(t),\ t \geq 0)$ have the law of the number of customers in an $M/M/1$ queue with arrival rate $\nu + \epsilon$ and service rate 1, where ϵ is taken to be small enough ensuring that the process Q_ϵ^+ is ergodic, i.e., that

$$\rho_\epsilon^+ := \nu + \epsilon < 1 \quad \text{for all } \epsilon \text{ small enough,} \tag{EC.1}$$

where ρ_ϵ^+ is the utilization of this $M/M/1$ queue. We add the subscript d to denote the initial value at time 0. That is, $Q^+_{\epsilon,d}(t)$ corresponds to the queue length at time t of the $M/M/1$ queue initialized (at time $t=0$) in state d. The initial condition can be random.

Note that D_I^λ is a Birth-and-Death (BD) process with death rate λ and birth rate in state d equal to $\mu_I(N_I^\lambda-(d-K_I^\lambda)^+)+\theta(K_I^\lambda-d)^+$. Since $\mu_I N_I^\lambda+\theta K_I^\lambda=\nu\lambda+o(\lambda)$ we have that

$$\mu_I(N_I^\lambda-(d-K_I^\lambda)^+)+\theta(K_I^\lambda-d)^+\le\lambda(\nu+\epsilon)=\lambda\rho_\epsilon^+ \tag{EC.2}$$

for all λ large enough and all $d\in\mathbb{Z}_+$. Hence, if we scale the birth (arrival) and death (service) rates of Q_ϵ^+ by λ, then both Q_ϵ^+ and D_I^λ have the same death rates, but the birth rates of Q_ϵ^+ are larger than those of D_I^λ. Scaling the rates of Q_ϵ^+ is tantamount to scaling its time argument by a factor of λ. We can thus prove the following ordering result using standard coupling arguments for BD processes (see, e.g., Lemma 1 in Whitt [1991]). The detailed proof is omitted. To simplify notation, we let

$$Y_d:=D_I^\lambda(0)\quad\text{and}\quad Y_q:=Q_\epsilon^+(0). \tag{EC.3}$$

LEMMA EC.1.1. (**upper bound for** D_I^λ) Fix $\epsilon>0$ and λ large enough so that (EC.2) holds, and assume that $Y_d\le_{st}Y_q$. Then,

$$\left\{D_I^\lambda(t),\ t\ge0\right\}\le_{st}\left\{Q^+_{\epsilon,Y_q}(\lambda t),\ t\ge0\right\}.$$

The above ordering allows us to upper bound the sequence D_I^λ by a single (time scaled) $M/M/1$ queue. The following auxiliary result is proved in §EC.1.3.

LEMMA EC.1.2. (**SSC in diffusion limit**) Suppose that (10) holds. Then, for all $\eta,T>0$,

$$\lim_{a\to\infty}\limsup_{\lambda\to\infty}\mathbb{P}\left\{\sup_{0\le s\le T}D_I^\lambda(s)\ge a\sqrt{\lambda}\right\}=0\quad\text{and}\quad\lim_{a\to\infty}\limsup_{\lambda\to\infty}\mathbb{P}\left\{\sup_{\eta\le s\le T}D_I^\lambda(s)\ge a\log\lambda\right\}=0.$$

Consequently,

$$\lambda^{-1/2}(Q_I^\lambda-K_I^\lambda)\Rightarrow0\quad\text{in }\mathcal{D}[0,\infty)\text{ as }\lambda\to\infty.$$

Note that the final conclusion of the lemma implies, in particular, the convergence $\widehat{X}_I^\lambda\Rightarrow0e$ in Theorem 4.1. We interpret this result as a state-space collapse result. It shows that, in the diffusion limit the state in

station I is constant so that the station O captures the state of the network. However, as explained in the introduction and further discussed in §4.2, this SSC result is not sufficient for our needs (see Example 1).

In addition to $Q^+_\epsilon(t)$, which serves as a sample-path stochastic-order upper bound for D^λ_I, we introduce a process, Q^-_ϵ, which corresponds to the queue-length process in a $M/M/1$ queue with service rate 1 and arrival rate $\nu-\epsilon$ and will serve as lower bound. The process Q^-_ϵ provides a weaker bound than the sample-path stochastic-order upper bound provided by Q^+_ϵ. That weaker bound is, however, sufficient for our needs. Below we choose ϵ sufficiently small so that both $\rho^+_\epsilon < 1$ (as defined in (EC.1)) and $\epsilon < \nu$. As before, we add a subscript to make explicit the dependency on the initial condition at time 0. Recall the definition of Y_d in (EC.3).

LEMMA EC.1.3. (**lower bound for** D^λ_I) Assume that (10) holds and that $Y^\lambda_d = O_P(\sqrt{\lambda})$. Then, given $T>0$, there exists a non-negative sequence ε^λ_T such that $\varepsilon^\lambda_T \to 0$ as $\lambda\to\infty$ and so that for all $d_1 \in \mathbb{Z}_+$,

$$\mathbb{P}\left\{D^\lambda_I(t) \geq d_1\right\} \geq \mathbb{P}\left\{Q^-_{\epsilon,Y^\lambda_d}(\lambda t) \geq d_1\right\} - \varepsilon^\lambda_T, \quad \text{for all } t\in[0,T].$$

Note that the conditions of the lemma are satisfied, in particular, if $Y^\lambda_d = b\sqrt{\lambda}$ for all λ and a non-random constant $b\geq 0$.

Since Q^-_ϵ and Q^+_ϵ have birth and death rates that do not scale with λ, one expects that $Q^-_{\epsilon,Y^\lambda_d}(\lambda t)$ and $Q^+_{\epsilon,Y^\lambda_d}(\lambda t)$ are close, in a sense, to their steady-state random variables $Q^+_\epsilon(\infty)$ and $Q^-_\epsilon(\infty)$ for all λ large enough and all $t>0$. Since these steady-state random variables are also "close" to each other (for small values of ϵ), one expects the same to hold for the process D^λ_I, i.e., that $D^\lambda_I(t)$ has approximately the distribution of $Q^-_\epsilon(\infty)$ (or $Q^+_\epsilon(\infty)$) for λ large enough and for all $t>0$. The following theorem formalizes this intuition in showing that $D^\lambda_I(t)$ converges to a local steady state instantaneously (pointwise, for each $t>0$) as $\lambda\to\infty$.

THEOREM EC.1.4. (**pointwise stationarity**) Fix a sequence $Y^\lambda_d = O_P(\sqrt{\lambda})$. Then, for all $t_1 > t_0$ and all $d_1\in\mathbb{Z}_+$,

$$\mathbb{P}\left\{D^\lambda_I(t_1)\geq d_1 \middle| D^\lambda_I(t_0) = Y^\lambda_d\right\} \Rightarrow \nu^{d_1} \quad \text{in } \mathbb{R} \text{ as } \lambda\to\infty. \tag{EC.4}$$

In particular, fixing $t_0 = 0$ and assuming (10), we have for all $t > 0$ that,

$$D_I^\lambda(t) \Rightarrow Q_b(\infty) \quad \text{in } \mathbb{R} \text{ as } \lambda \to \infty, \tag{EC.5}$$

where $\mathbb{P}\{Q_b(\infty) \geq d_1\} = \nu^{d_1}$.

Note that the convergence in (EC.4) is convergence in distribution, since, for each fixed λ, the conditional probabilities are random variables rather than numbers.

To prove Theorem EC.1.4 we will need the following lemma, whose proof appears in §EC.1.3.

LEMMA EC.1.5. Let $Y_d^\lambda = O_p(\sqrt{\lambda})$. Then, for any $t > 0$,

$$\lim_{\lambda\to\infty} \mathbb{P}\left\{Q^+_{\epsilon,Y_d^\lambda}(\lambda t) \geq d_1\right\} = \mathbb{P}\left\{Q^+_\epsilon(\infty) \geq d_1\right\} \text{ and } \lim_{\lambda\to\infty} \mathbb{P}\left\{Q^-_{\epsilon,Y_d^\lambda}(\lambda t) \geq d_1\right\} = \mathbb{P}\left\{Q^-_\epsilon(\infty) \geq d_1\right\}.$$

Proof of Theorem EC.1.4: Since the process D_I^λ is Markovian it suffices to prove (EC.4) for $t_0 = 0$ and $t := t_1 - t_0$. We start by fixing $\delta > 0$. We will show that, given $\varepsilon, \delta > 0$, it holds for all sufficiently large λ, that

$$\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t) \geq d_1 \middle| D_I^\lambda(0) = Y_d^\lambda\right\} \geq \mathbb{P}\left\{Q^-_\epsilon(\infty) \geq d_1\right\} - \delta\right\} \geq 1 - \varepsilon, \tag{EC.6}$$

and

$$\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t) \geq d_1 \middle| D_I^\lambda(0) = Y_d^\lambda\right\} \leq \mathbb{P}\left\{Q^+_\epsilon(\infty) \geq d_1\right\} + \delta\right\} \geq 1 - \varepsilon. \tag{EC.7}$$

The steady-state distribution of Q^+_ϵ is continuous in ϵ in the sense that $Q^+_\epsilon(\infty) \Rightarrow Q_b(\infty)$ and $Q^-_\epsilon(\infty) \Rightarrow Q_b(\infty)$ as $\epsilon \downarrow 0$, and where $Q_b(\infty)$ has the steady-state distribution of a $M/M/1$ with service rate 1 and arrival rate ν; see e.g. Lemma 9.8 in Perry and Whitt [2010b] (where the statement is proved for a more general quasi-birth-and-death (QBD) process. Note that α denotes the steady state distribution in that reference). In particular, for all sufficiently small ϵ

$$\left|\mathbb{P}\left\{Q^+_\epsilon(\infty) \geq d_1\right\} - \mathbb{P}\left\{Q_b(\infty) \geq d_1\right\}\right| \leq \delta \quad \text{and} \quad \left|\mathbb{P}\left\{Q^-_\epsilon(\infty) \geq d_1\right\} - \mathbb{P}\left\{Q_b(\infty) \geq d_1\right\}\right| \leq \delta.$$

Plugging this into (EC.6) and (EC.7) we then have that

$$\liminf_{\lambda\to\infty} \mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t) \geq d_1 \middle| D_I^\lambda(0) = Y_d^\lambda\right\} \geq \mathbb{P}\left\{Q_b(\infty) \geq d_1\right\} - 2\delta\right\} \geq 1 - \varepsilon$$

and

$$\liminf_{\lambda\to\infty} \mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=Y_d^\lambda\right\}\leq \mathbb{P}\left\{Q_b(\infty)\geq d_1\right\}+2\delta\right\}\geq 1-\varepsilon.$$

Since δ and ε are arbitrary we may conclude that $\mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=Y_d^\lambda\right\}$ converges in probability (and thus in distribution) to the constant $\mathbb{P}\left\{Q_b(\infty)\geq d_1\right\}$ and, because the distribution is discrete, the convergence is, in fact, uniform over compact subsets of $\mathbb{Z}_+$. In passing we note that the arguments to prove (EC.6) and (EC.7) can be repeated for any sequence of initial conditions $Y_d^\lambda = O(\sqrt{\lambda})$.

The remainder of the proof is dedicated to establishing (EC.6) and (EC.7), starting with the former. Note that

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=Y_d^\lambda\right\}\geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta\right\}\\
&\quad=\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=Y_d^\lambda\right\}\geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta;\ Y_d^\lambda > b\sqrt{\lambda}\right\}\\
&\qquad+\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=Y_d^\lambda\right\}\geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta;\ Y_d^\lambda \leq b\sqrt{\lambda}\right\}\\
&\quad\geq\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=Y_d^\lambda\right\}\geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta;\ Y_d^\lambda \leq b\sqrt{\lambda}\right\}.
\end{aligned}$$

From the monotonicity of D_I^λ in its initial condition, see, e.g., Chapter 9 in Ross [1996], we have that almost surely

$$\mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=Y_d^\lambda\right\}\geq \mathbb{P}\left\{D_I^\lambda(t)\geq d_1 \middle| D_I^\lambda(0)=0\right\}, \tag{EC.8}$$

and we note that, while the LHS of the inequality is a random variable, the probability on the RHS is a constant. We thus have

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{P}\{D_I^\lambda(t)\geq d_1 \mid D_I^\lambda(0)=Y_d^\lambda\}\geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta;\ Y_d^\lambda \leq b\sqrt{\lambda}\right\} \qquad \text{(EC.9)}\\
&=\mathbb{P}\left\{\mathbb{P}\{D_I^\lambda(t)\geq d_1 \mid D_I^\lambda(0)=Y_d^\lambda\}\geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta \;\middle|\; Y_d^\lambda \leq b\sqrt{\lambda}\right\}\mathbb{P}\left\{Y_d^\lambda \leq b\sqrt{\lambda}\right\}\\
&\geq \mathbb{1}\left\{\mathbb{P}\{D_I^\lambda(t)\geq d_1 \mid D_I^\lambda(0)=0\}\geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta\right\}\mathbb{P}\left\{Y_d^\lambda \leq b\sqrt{\lambda}\right\}\\
&\geq \mathbb{1}\left\{\mathbb{P}\{Q_{\epsilon,0}^-(\lambda t)\geq d_1\}-\varepsilon_T^\lambda \geq \mathbb{P}\left\{Q_\epsilon^-(\infty)\geq d_1\right\}-\delta\right\}\mathbb{P}\left\{Y_d^\lambda \leq b\sqrt{\lambda}\right\}.
\end{aligned}$$

Here, the first inequality follows from (EC.8) and noting that, conditioned on $D_I^\lambda(0)=0$, the conditional probability becomes a constant, specifically, it is equal to either 0 or 1, so that it can be replaced

with the indicator. Using Lemma EC.1.5 we have, for all λ large enough, and for a fixed $\delta > 0$, that $\mathbb{P}\left\{Q^-_{\epsilon,0}(\lambda t) \geq d_1\right\} - \varepsilon^\lambda_T \geq \mathbb{P}\left\{Q^-_\epsilon(\infty) \geq d_1\right\} - \delta$ (because $\varepsilon^\lambda_T \to 0$ as $\lambda \to \infty$), so that the indicator in (EC.9) is in fact 1 for all λ large enough. Also, by the assumption of the theorem, there exists $b > 0$ such that for any given $\varepsilon > 0$ and for all λ large enough, $\mathbb{P}\left\{Y^\lambda_d \leq b\sqrt{\lambda}\right\} \geq 1 - \varepsilon$. Hence, it follows from (EC.9) that for all λ large enough,

$$\mathbb{P}\left\{\mathbb{P}\left\{D^\lambda_I(t) \geq d_1 \middle| D^\lambda_I(0) = Y^\lambda_d\right\} \geq \mathbb{P}\left\{Q^-_\epsilon(\infty) \geq d_1\right\} - \delta\right\} \geq 1 - \varepsilon,$$

This establishes (EC.6).

The arguments for establishing (EC.7) are very similar. We replace (EC.9) with

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{P}\{D^\lambda_I(t) \geq d_1 \mid D^\lambda_I(0) = Y^\lambda_d\} \leq \mathbb{P}\left\{Q^+_\epsilon(\infty) \geq d_1\right\} + \delta;\ Y^\lambda_d \leq b\sqrt{\lambda}\right\} \\
&\quad \geq \mathbb{1}\left\{\mathbb{P}\{D^\lambda_I(t) \geq d_1 \mid D^\lambda_I(0) = b\sqrt{\lambda}\} \leq \mathbb{P}\left\{Q^-_\epsilon(\infty) \geq d_1\right\} + \delta\right\} \mathbb{P}\left\{Y^\lambda_d \leq b\sqrt{\lambda}\right\} \\
&\quad \geq \mathbb{1}\left\{\mathbb{P}\{Q^+_{\epsilon,b\sqrt{\lambda}}(\lambda t) \geq d_1\} \leq \mathbb{P}\left\{Q^-_\epsilon(\infty) \geq d_1\right\} + \delta\right\} \mathbb{P}\left\{Y^\lambda_d \leq b\sqrt{\lambda}\right\}.
\end{aligned}$$

Here, the first inequality follows again from the monotonicity of the process D^λ_I and from the fact that, conditioned on $D^\lambda_I(0)$ being equal to a constant, the conditional probabilities become constants. The last inequality above follows from Lemma EC.1.1. Together with Lemma EC.1.5, (EC.7) is established. ■

Next, we establish stochastic-process limits for the cumulative (in)availability process as defined in equation (9). We define the related scaled and centered process

$$\widehat{C}^\lambda_I(t) = \sqrt{\lambda}\left(C^\lambda_I(t) - \frac{(\lambda - \mu_I N^\lambda_I - \theta K^\lambda_I)}{\lambda} t\right), \quad t \geq 0.$$

Theorem EC.1.6 below shows that this scaling leads to meaningful limits, by establishing a FCLT for $\widehat{C}^\lambda_I$ via the AP discussed in Remark 4.2. We note that, even though the AP is related to the pointwise stationarity in Theorem EC.1.4, it is not directly implied by it (nor does the AP imply Theorem EC.1.4). Both phenomena are closely related being consequences of the separation of time scales that is caused by the fast oscillations of D^λ_I.

THEOREM EC.1.6. (**FCLT via the AP**) Suppose that the conditions of Theorem 4.1 hold. Then $\widehat{C}^\lambda_I \Rightarrow \check{\sigma} B$ in $\mathcal{D}[0,\infty)$ as $\lambda \to \infty$, where B is a standard Brownian motion and $\check{\sigma}^2 = 2\nu$.

The fact that (time-scaled) cumulative processes associated with regenerative processes converge to Brownian limits under proper scalings of time and space is not new; see Glynn and Whitt [1993]. The important thing to observe here is that, in contrast to the results in Glynn and Whitt [1993], we do not scale time to get the Brownian limit. As the proof reveals, this is due to the fast oscillations of the process D_I^λ, which completes $O(\lambda)$ (regenerative) cycles over any finite time interval. See also Perry and Whitt [2010c].

The proof of Theorem EC.1.6 builds on two steps: (i) transformation of the fast oscillations of the process D_I^λ over intervals of length $[0,T)$ to oscillations of a related "slowed-down" process $D_I^{s,\lambda}$ over intervals of length $[0,\lambda T)$ (see (EC.10) below), and (ii) applying arguments in the spirit of Glynn and Whitt [1993] in the proof of the FCLT for the "slower" process $D_I^{s,\lambda}$.

We first define the "slowed down" process $D_I^{s,\lambda}$. To that end, recall that the process D_I^λ has the probability law of a single-server queue with state-dependent arrival rates. Specifically, the service rate is λ, while the arrival rate is $\mu(N_I^\lambda-(d-K_I^\lambda)^+)+\theta(K_I^\lambda-d)^+$ when in state $d\in\left\{0,1,\ldots,N_I^\lambda+K_I^\lambda\right\}$. We then define the "slowed down" processes for each λ

$$D_I^{s,\lambda}(t):=D_I^\lambda(t/\lambda),\quad t\geq 0. \tag{EC.10}$$

Then $D_I^{s,\lambda}$ has the probability law of a state-dependent $M/M/1$ queue with service rate $\lambda/\lambda=1$ and state-dependent arrival rate $\lambda^{-1}(\mu(N_I^\lambda-(d-K_I^\lambda)^+)+\theta(K_I^\lambda-d)^+)$. Moreover,

$$C_I^\lambda(t):=\int_0^t \mathbb{1}\left\{D_I^\lambda(s)=0\right\}ds=\frac{1}{\lambda}\int_0^{\lambda t}\mathbb{1}\left\{D_I^{s,\lambda}(s)=0\right\}ds,\quad t\geq 0.$$

By Assumption 1, the arrival rate of $D_I^{s,\lambda}$ is bounded by $(\mu_I N_I^\lambda+\theta K_I^\lambda)/\lambda=\nu+o(1)$ where $\nu<1$ so that $D_I^{s,\lambda}$ is (for all λ large enough) a stable queue. We define ρ_d^λ to be the steady-state probability that the server is busy in this state-dependent $M/M/1$ queue.

The following theorem restates Theorem EC.1.6 in terms of the slowed-down processes $D_I^{s,\lambda}$.

THEOREM EC.1.7. Under the conditions of Theorem 4.1,

$$\sqrt{\lambda}\left(\frac{1}{\lambda}\int_0^{\lambda\cdot}\mathbb{1}\left\{D_I^{s,\lambda}(u)=0\right\}du-(1-\rho_d^\lambda)e\right)\Rightarrow\tilde{\sigma}B\quad\text{in }\mathcal{D}[0,\infty)\text{ as }\lambda\to\infty,$$

where $\tilde{\sigma}^2=2\nu$. Furthermore, $\sqrt{\lambda}\left((1-\rho_d^\lambda)-(\lambda-\mu_I N_I^\lambda-\theta K_I^\lambda)/\lambda\right)\to 0$ as $\lambda\to\infty$, so that

$$\sqrt{\lambda}\left(\frac{1}{\lambda}\int_0^{\lambda\cdot}\mathbb{1}\left\{D_I^{s,\lambda}(u)=0\right\}du-\frac{\lambda-\mu_I N_I^\lambda-\theta K_I^\lambda}{\lambda}e\right)\Rightarrow\tilde{\sigma}B\quad\text{in }\mathcal{D}[0,\infty)\text{ as }\lambda\to\infty.$$

Proof: To formally define the underlying regenerative process let $\tau^\lambda_{-1} = 0$, and for $k \geq 0$, define recursively

$$\tau^\lambda_k := \inf\left\{u \geq \tau^\lambda_{k-1} : D^{s,\lambda}_I(u-) \neq 0,\ D^{s,\lambda}_I(u) = 0\right\}.$$

For $k \geq 0$, define

$$\Psi^\lambda_k := \int_{\tau^\lambda_{k-1}}^{\tau^\lambda_k} \left(\mathbb{1}\left\{D^{s,\lambda}_I(u) = 0\right\} - (1-\rho^\lambda_d)\right) du.$$

For $k \geq 1$, $T^\lambda_k := \tau^\lambda_k - \tau^\lambda_{k-1}$ is then the length of the k^{th} busy cycle (the busy cycle consists of the busy period and the idle period). Since $\left\{T^\lambda_k,\ k \geq 1\right\}$ are IID, we will use T^λ_1 to denote a general random variable with the distribution of a busy cycle.

Let $\left\{\xi^\lambda_k,\ k \geq 1\right\}$ be IID exponential random variables with rate $(\mu_I N^\lambda_I + \theta K^\lambda_I)/\lambda$, corresponding to the time spent at state 0 during a busy cycle and define $\bar{T}^\lambda_k := \tau^\lambda_k - \tau^\lambda_{k-1} - \xi^\lambda_k$. Then $\left\{\bar{T}^\lambda_k,\ k \geq 1\right\}$ are IID with each having the distribution of the busy period of this state-dependent $M/M/1$ queue. We can thus write

$$\Psi^\lambda_k = \xi^\lambda_k \rho^\lambda_d - (1-\rho^\lambda_d)\bar{T}^\lambda_k, \quad k \geq 1, \tag{EC.11}$$

where $\Psi^\lambda_0 = 0$ by definition. We define the corresponding (possibly delayed) renewal process

$$R^\lambda(t) := \sup\left\{k \geq 0 : \tau^\lambda_k \leq t\right\}.$$

Since $\Psi^\lambda_0 = 0$, we can write

$$\frac{1}{\lambda}\int_0^{\lambda t} \mathbb{1}\left\{D^{s,\lambda}_I(u) = 0\right\} du - (1-\rho^\lambda_d)t = \sum_{k=1}^{R^\lambda(\lambda t)} \frac{\Psi^\lambda_k}{\lambda} + \frac{1}{\lambda}\int_{\tau_{R^\lambda(\lambda t)}}^{\lambda t} \left(\mathbb{1}\left\{D^{s,\lambda}_I(u) = 0\right\} - (1-\rho^\lambda_d)\right) du. \tag{EC.12}$$

From here the argument follows very closely the standard proof of the FCLT for regenerative processes. Some care is needed because the IID random variables $\left\{\Psi^\lambda_k\right\}_{k\geq 1}$ are indexed by both k and λ. This entails replacing some parts of the standard proof with an FCLT for triangular arrays. It also entails establishing bounds to identify the asymptotic variance terms.

First, using (EC.11), we write

$$\sqrt{\lambda}\sum_{k=1}^{R^\lambda(\lambda t)} \frac{\Psi^\lambda_k}{\lambda} = \frac{1}{\sqrt{\lambda}}\sum_{k=1}^{R^\lambda(\lambda t)} \rho^\lambda_d \xi^\lambda_k - \frac{1}{\sqrt{\lambda}}\sum_{k=1}^{R^\lambda(\lambda t)} (1-\rho^\lambda_d)\bar{T}^\lambda_k$$

$$= \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} \left(\rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} \right) - \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} (1-\rho_d^\lambda) \left(\bar{T}_k^\lambda - \mathbb{E}\left[\bar{T}_k^\lambda\right] \right)$$
$$+ \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} \left(\frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} - (1-\rho_d^\lambda) \mathbb{E}\left[\bar{T}_k^\lambda\right] \right). \tag{EC.13}$$

The next two lemmas identify asymptotic expressions of the expectation and variance terms of the cycle-related random variables. These expressions are applied to establish an FCLT for the partial sums in (EC.13). The proofs of Lemmas EC.1.8 and EC.1.9 appear in §EC.1.3.

LEMMA EC.1.8. As $\lambda \to \infty$,

$$\left(\sum_{k=1}^{\lceil \lambda \cdot \rceil} Var\left(\frac{\xi_k^\lambda}{\sqrt{\lambda}} \right), \sum_{k=1}^{\lceil \lambda \cdot \rceil} Var\left(\frac{\bar{T}_k^\lambda}{\sqrt{\lambda}} \right), \frac{R^\lambda(\lambda \cdot)}{\lambda} \right) \Rightarrow \left(\frac{1}{\nu^2} e, \frac{1+\nu}{(1-\nu)^3} e, \nu(1-\nu) e \right) \quad \text{in } \mathcal{D}[0,\infty). \tag{EC.14}$$

Consequently, as $\lambda \to \infty$,

$$\left(\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda \cdot)} \left(\rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} \right), \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda \cdot)} (1-\rho_d^\lambda) \left(\bar{T}_k^\lambda - \mathbb{E}\left[\bar{T}_k^\lambda\right] \right) \right) \Rightarrow \tag{EC.15}$$
$$\left(\sqrt{\nu(1-\nu)} B^1, \sqrt{\nu(1+\nu)} B^2 \right)$$

in $D^2[0,\infty)$, where B^1 and B^2 are two independent standard Brownian motions.

LEMMA EC.1.9. Under the conditions of Theorem 4.1,

$$\sqrt{\lambda} \left(\frac{\rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} - (1-\rho_d^\lambda) \mathbb{E}\left[\bar{T}_1^\lambda\right] \right) \to 0 \quad \text{as} \quad \lambda \to \infty.$$

In turn,

$$\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda \cdot)} \left(\frac{\rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} - (1-\rho_d^\lambda) \mathbb{E}\left[\bar{T}_1^\lambda\right] \right) \Rightarrow 0 \quad \text{in } D[0,\infty) \text{ as } \lambda \to \infty. \tag{EC.16}$$

Proceeding with the proof of the theorem, if

$$\frac{1}{\lambda} \int_{\tau_{R^\lambda(\lambda t)}}^{\lambda t} \left(\mathbb{1}\left\{ D_I^{s,\lambda}(u) = 0 \right\} - (1-\rho_d^\lambda) \right) du \Rightarrow 0 \quad \text{as} \quad \lambda \to \infty, \tag{EC.17}$$

then the theorem is established by replacing the sum in (EC.12) with its version in (EC.13) and then replacing the right-hand side (RHS) in (EC.13) with (EC.15) and (EC.16), and noting the sum of the variance of

the Brownian motions B^1 and B^2 in (EC.15) is $\tilde{\sigma}^2 = 2\nu$. It thus remains only to establish (EC.17). We first note that for all $u \geq 0$, $|\mathbb{1}\left\{D_I^{s,\lambda}(u) = 0\right\} - (1 - \rho_d^\lambda)| \leq 2$ so that, to show (EC.17), it suffices to show that

$$\frac{2}{\sqrt{\lambda}}\left(\tau^\lambda_{R^\lambda(\lambda t)+1} - \tau^\lambda_{R^\lambda(\lambda t)}\right) \Rightarrow 0 \quad \text{as} \quad \lambda \to \infty.$$

To that end, from the convergence of the variance terms in Lemma EC.1.8, it follows that $\mathbb{E}\left[(T_k^\lambda)^2\right] \leq C$ for all λ large enough and a constant C that does not depend of λ. By Markov's inequality, $\mathbb{P}\left\{T_k^\lambda > \epsilon\sqrt{\lambda}\right\} \leq C/(\epsilon^2\lambda)$ for all λ large enough. In turn, for any constant $M > 0$,

$$\lambda^{-1/2} \max_{1 \leq k \leq \lambda M} T_k^\lambda \Rightarrow 0 \quad \text{as } \lambda \to \infty. \tag{EC.18}$$

(See e.g. the proof of Lemma 2.3.1 in the online appendix to Whitt [2002a].) It follows from (EC.14) and (EC.18) that for any fixed $T > 0$ and for M_T to be specified shortly,

$$\mathbb{P}\left\{\sup_{0\leq t\leq T} \frac{1}{\sqrt{\lambda}} \max_{1\leq k\leq R^\lambda(\lambda t)+1} T_k^\lambda > \epsilon\right\} \leq \mathbb{P}\left\{\frac{1}{\sqrt{\lambda}} \max_{1\leq k\leq \lambda M_T} T_k^\lambda > \epsilon\right\} + \mathbb{P}\left\{\sup_{0\leq t\leq T} R^\lambda(\lambda t) + 1 > \lambda M_T\right\}.$$

Indeed, the first element in the RHS above converges to 0 by (EC.18), and the second element converges to 0 by (EC.14), provided that $M_T > \nu(1-\nu)T$. Hence $T^\lambda_{R^\lambda(\lambda t)+1}/\sqrt{\lambda} \Rightarrow 0$ in $\mathcal{D}[0,\infty)$ as $\lambda \to \infty$. This concludes the proof of the theorem. ■

EC.1.2. Proofs of Main Results

In this section we prove the main results stated in §4 in the order of their appearance.

Proof of Theorem 4.1: The overflow process is the number of arrivals by time t that find X_I^λ equal to $N_I + K_I$ upon arrival or, equivalently, find $D_I^\lambda = 0$ when they arrive. Let t_k^λ be the arrival time of the k^{th} customer to arrive (the k^{th} jump of the exogenous arrival process $A^\lambda(t)$). Then, $A_O^\lambda(t) = \sum_{k=1}^{A^\lambda(t)} 1\{D_I^\lambda(t_k^\lambda -) = 0\}$ or, in simpler form, $A_O^\lambda(t) = \int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} dA^\lambda(s)$. In turn, the scaled process (see §3.2) satisfies

$$\begin{aligned}\widehat{A}_O^\lambda(t) = &\frac{\int_0^t \mathbb{1}\left\{D_I^\lambda(s-) = 0\right\}\, dA^\lambda(s) - \lambda\int_0^t \mathbb{1}\left\{D_I^\lambda(s-) = 0\right\}\, ds}{\sqrt{\lambda}} \\ &+ \frac{\lambda\int_0^t \mathbb{1}\left\{D_I^\lambda(s-) = 0\right\}\, ds - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)t}{\sqrt{\lambda}}\end{aligned} \tag{EC.19}$$

We treat each of the elements on the RHS of (EC.19) separately. For the first element define

$$\widehat{M}^\lambda(t) := \frac{1}{\sqrt{\lambda}}\left(\int_0^t \mathbb{1}\left\{D_I^\lambda(s-) = 0\right\}\, dA^\lambda(s) - \lambda\int_0^t \mathbb{1}\left\{D_I^\lambda(s-) = 0\right\} ds\right)$$

$$= \frac{1}{\sqrt{\lambda}} \left(\int_0^t \mathbb{1}\left\{ D_I^\lambda(s-) = 0 \right\} \, d(A^\lambda(s) - \lambda s) \right).$$

Note that the process $\widehat{M}^\lambda(t)$ is a square integrable martingale with respect to the filtration $\mathcal{F}^\lambda = (\mathcal{F}_t^\lambda)_{t\geq 0}$, where $\mathcal{F}_t^\lambda = \sigma\left\{ (D_I^\lambda(s), Q_O^\lambda(s), A^\lambda(s));\ s \leq t \right\}$; having a predictable quadratic variation process $\langle \widehat{M}^\lambda \rangle(t) = \int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} ds$; see, e.g., Lemma 3.2 in Pang et al. [2007]. By Theorem EC.1.6 we have that

$$\frac{\lambda \int_0^\cdot \mathbb{1}\left\{ D_I^\lambda(s-) = 0 \right\} ds - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda) e}{\lambda} \Rightarrow 0 \quad \text{in } \mathcal{D}[0,\infty) \text{ as } \lambda \to \infty.$$

By Assumption 1, $(\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)/\lambda \to (1-\nu)$ as $\lambda \to \infty$, so that

$$\langle \widehat{M}^\lambda \rangle(\cdot) = \int_0^\cdot \mathbb{1}\left\{ D_I^\lambda(s-) = 0 \right\} ds \Rightarrow (1-\nu) e \quad \text{in } \mathcal{D}[0,\infty) \text{ as } \lambda \to \infty.$$

From here it follows by the Martingale FCLT (see, e.g., Theorem 8.1 in Pang et al. [2007]) that

$$\frac{\int_0^\cdot \mathbb{1}\left\{ D_I^\lambda(s-) = 0 \right\} \, dA^\lambda(s) - \lambda \int_0^\cdot \mathbb{1}\left\{ D_I^\lambda(s-) = 0 \right\} \, ds}{\sqrt{\lambda}} \Rightarrow \sqrt{(1-\nu)} B^1 \quad \text{in } \mathcal{D}[0,\infty) \text{ as } \lambda \to \infty.$$

For the second element on the RHS of (EC.19) we have by Theorem EC.1.6, that

$$\frac{\lambda \int_0^\cdot \mathbb{1}\left\{ D_I^\lambda(s) = 0 \right\} \, ds - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda) e}{\sqrt{\lambda}} \Rightarrow \sqrt{2\nu} B^2 \quad \text{in } \mathcal{D}[0,\infty) \text{ as } \lambda \to \infty.$$

The convergence of $\widehat{A}_O^\lambda$ now follows from these last two limits and from the continuity of the addition mapping at continuous limits.

Finally, by Lemma EC.1.2 we have that $\widehat{X}_I^\lambda \Rightarrow 0$ in $\mathcal{D}(0,\infty)$ as $\lambda \to \infty$. The marginal convergence of $\widehat{A}_O^\lambda$, $\widehat{X}_I^\lambda$ and $\langle \widehat{M}^\lambda \rangle$ now implies their joint convergence because $\widehat{X}_I^\lambda$ and $\langle \widehat{M}^\lambda \rangle$ have deterministic limits; see e.g. Theorem 11.4.5 in Whitt [2002a]. ■

Proof of Corollary 4.2: Recall that $X_I^\lambda(\infty)$ has the distribution of the steady-state number of customers in an $M/M/N_I^\lambda/K_I^\lambda + M$ queue with arrival rate λ, N_I^λ servers, service rate μ and waiting room of size K_I^λ. Also

$$\widehat{X}_I^\lambda(t) = \frac{X_I^\lambda(t) - (N_I^\lambda - K_I^\lambda)}{\sqrt{\lambda}} = -\frac{D_I^\lambda(t)}{\sqrt{\lambda}} \geq 0.$$

We will show that

$$\mathbb{E}[|\widehat{X}_I^\lambda(\infty)|] \to 0 \qquad \text{as} \quad \lambda \to \infty, \tag{EC.20}$$

so that, in particular, $\widehat{X}_I^\lambda(\infty) \Rightarrow 0$. This would imply that, initializing station I at time $t=0$ with its steady-state distribution, $\{\widehat{X}_I^\lambda(0)\}$ satisfies (10) so that the convergence of the scaled overflow process follows from Theorem 4.1. In fact, (11) also follows from (EC.20). Indeed, since inflow=outflow in steady-state we have $\lambda(1-p_b^\lambda) = \mu\mathbb{E}[Z_I^\lambda(\infty)] + \theta\mathbb{E}[Q^\lambda(\infty)]$ where $Z_I^\lambda(\infty)$ has the steady-state distribution of the number of busy servers in station I. By work conservation in station I we have that $Z_I^\lambda(\infty) = X_I^\lambda(\infty) \wedge N_I^\lambda$ and $Q^\lambda(\infty) = (X_I^\lambda(\infty) - N_I^\lambda)^+$. Then, by (EC.20),

$$\mathbb{E}[Z_I^\lambda(\infty)] = N_I^\lambda - \mathbb{E}[(X_I^\lambda(\infty) - N_I^\lambda)^-] = N_I^\lambda - o(\sqrt{\lambda})$$

and

$$\mathbb{E}[Q^\lambda(\infty)] = \mathbb{E}[(X_I^\lambda(\infty) - N_I^\lambda)^+] = K_I^\lambda + o(\sqrt{\lambda})$$

and, in turn, that $\lambda(1-p_b^\lambda) = \mu_I N_I^\lambda + \theta K_I^\lambda + o_p(\sqrt{\lambda})$. Equation (11) is now obtained by dividing by λ.

To conclude the proof it remains to prove (EC.20). This, we claim, follows immediately from Lemma EC.1.1. Indeed, from that lemma we have that for every $\epsilon > 0$ we can find λ large enough, so that $D_I^\lambda(\infty) \leq_{st} Q_\epsilon^+(\infty)$, where $Q_\epsilon^+(\infty)$ has the steady-state distribution of an $M/M/1$ queue with utilization $\rho_\epsilon^+ < 1$, for ρ_ϵ^+ in (EC.1). Hence,

$$\mathbb{E}\left[D_I^\lambda(\infty)\right] \leq_{st} \mathbb{E}\left[Q_\epsilon^+(\infty)\right] = O_p(1) = o_P(\sqrt{\lambda}).$$

In turn, after dividing by $\sqrt{\lambda}$ we get $\mathbb{E}\left[|\widehat{X}_I^\lambda(\infty)|\right] = \mathbb{E}\left[D_I^\lambda(\infty)/\sqrt{\lambda}\right] = o(1)$. This concludes the proof. ■

Proof of Theorem 4.3 Define for every λ the filtration $\mathcal{F}^\lambda = (\mathcal{F}_t^\lambda)_{t\geq 0}$ by

$$\mathcal{F}_t^\lambda = \sigma\left\{D_I^\lambda(s), X_O^\lambda(s); \quad s \leq t\right\},$$

and consider the filtered probability space $(\Omega, \mathbb{F}^\lambda, (\mathcal{F}_t^\lambda)_{t\geq 0}, \mathbb{P})$ (where $\mathbb{F}^\lambda$ is the σ-algebra over Ω, and $\mathcal{F}_t^\lambda \subset \mathbb{F}^\lambda$ for all $t \geq 0$).

First we claim that, from the pointwise stationarity of D_I^λ (see Theorem EC.1.4), it follows that for each $t > 0$ and $\delta \in (0,t)$, and $d \in \mathbb{Z}_+$,

$$\mathbb{P}\left\{D_I^\lambda(t) = d \middle| \mathcal{F}_{t-\delta}^\lambda\right\} \Rightarrow \mathbb{P}\left\{Q_b(\infty) = d\right\}, \text{ as } \lambda \to \infty. \tag{EC.21}$$

Since the conditional probability is bounded by 1, we have by dominated convergence that

$$\mathbb{E}\left[\mathbb{P}\left\{D_I^\lambda(t)=d\middle|\mathcal{F}_{t-\delta}^\lambda\right\}\right]\to\mathbb{P}\left\{Q_b(\infty)=d\right\},\text{ as }\lambda\to\infty. \tag{EC.22}$$

In words, $D_I^\lambda(t)$ is asymptotically independent of $\mathcal{F}_{t-\delta}^\lambda$. Indeed, for all $0\le s<t$, $D_I^\lambda(t)-D_I^\lambda(s)$ depends only on the arrivals on the interval $(s,t]$, the abandonments and the service completions on that interval, all of which, conditioned on $D_I^\lambda(s)$, are independent of $\mathcal{F}_s$. Note also that D_I^λ is Markov, so that $\mathbb{E}\left[\mathbb{1}\left\{D_I^\lambda(t)=d\right\}\middle|\mathcal{F}_{t-\delta}^\lambda\right]=\mathbb{E}\left[\mathbb{1}\left\{D_I^\lambda(t)=d\right\}\middle|D_I^\lambda(t-\delta)\right]$. By Lemma EC.1.2, $\sup_{0\le s\le t-\delta}D_I^\lambda(u)=O_P(\sqrt{\lambda})$, so that the (EC.21) follows from Theorem EC.1.4.

We next show how the statement of Theorem 4.3 follows from (EC.21). For all strictly positive T,ϵ and δ define

$$E_T:=E_T(\lambda,\delta,\epsilon)=\left\{\omega\in\Omega:\sup_{s,t\le T:|t-s|\le\delta}|\widehat{X}_O^\lambda(t)-\widehat{X}_O^\lambda(s)|\le\epsilon\right\}.$$

We denote by E_T^c the complement of E_T^c, i.e., $E_T^c:=\{\omega\in\Omega:\omega\notin E_T\}$. (We omit the dependency of $\widehat{X}_O^\lambda$ on ω to simplify notation, with the understanding that $\widehat{X}_O^\lambda(t)=\widehat{X}_O^\lambda(t,\omega)$.) From the $\mathcal{C}$-Tightness of $\widehat{X}_O^\lambda$ (see e.g. Theorem 3.1 of Whitt [2005] for the convergence of the underlying sequence of $GI/M/N_O^\lambda+M$ queues to a continuous limit), it follows that

$$\lim_{\delta\to0}\liminf_{\lambda\to\infty}\mathbb{P}\left\{E_T(\lambda,\delta,\epsilon)\right\}=1. \tag{EC.23}$$

We write

$$\begin{aligned}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q,D_I^\lambda(t)=d\right\}=&\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q,D_I^\lambda(t)=d,\mathbb{1}\left\{E_T\right\}\right\}\\&+\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q,D_I^\lambda(t)=d,\mathbb{1}\left\{E_T^c\right\}\right\}.\end{aligned} \tag{EC.24}$$

From (EC.23) it then follows that

$$\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q,D_I^\lambda(t)=d\right\}=\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q,D_I^\lambda(t)=d,\mathbb{1}\left\{E_T\right\}\right\}+o(1). \tag{EC.25}$$

Note that for all $\omega\in E_T$, $t\le T$ and $q\in\mathbb{R}^+$

$$\left\{\widehat{X}_O^\lambda(t-\delta)>q-\epsilon\right\}\subset\left\{\widehat{X}_O^\lambda(t)>q\right\}. \tag{EC.26}$$

By the same argument that leads to (EC.25) we have that

$$\mathbb{P}\left\{\widehat{X}_O^\lambda(t-\delta)>q-\epsilon, D_I^\lambda(t)=d, \mathbb{1}\left\{E_T\right\}\right\}=\mathbb{P}\left\{\widehat{X}_O^\lambda(t-\delta)>q-\epsilon, D_I^\lambda(t)=d\right\}+o(1),$$

so that, by (EC.26),

$$\begin{aligned}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q, D_I^\lambda(t)=d, \mathbb{1}\left\{E_T\right\}\right\} &\geq \mathbb{P}\left\{\widehat{X}_O^\lambda(t-\delta)>q-\epsilon, D_I^\lambda(t)=d\right\}+o(1)\\ &=\mathbb{E}\mathbb{E}\left[\mathbb{1}\left\{\widehat{X}_O^\lambda(t-\delta)>q-\epsilon\right\}\mathbb{1}\left\{D_I^\lambda(t)=d\right\}\middle|\mathcal{F}_{t-\delta}^\lambda\right]+o(1)\\ &=\mathbb{E}\left[\mathbb{1}\left\{\widehat{X}_O^\lambda(t-\delta)>q-\epsilon\right\}\right]\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{D_I^\lambda(t)=d\right\}\middle|\mathcal{F}_{t-\delta}^\lambda\right]\right]+o(1).\end{aligned}$$

The last equality above follows from the fact that, $\widehat{X}_O^\lambda(t-\delta)$ is measurable with respect to $\mathcal{F}_{t-\delta}^\lambda$. Using (EC.25) and applying (EC.22), we have that for all $\epsilon>0$,

$$\begin{aligned}\liminf_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q, D_I^\lambda(t)=d\right\} &\geq \lim_{\delta\to 0}\liminf_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t-\delta)>q-\epsilon\right\}\mathbb{E}\left[\mathbb{P}\left\{D_I^\lambda(t)=d\middle|\mathcal{F}_{t-\delta}^\lambda\right\}\right]\\ &=\lim_{\delta\to 0}\mathbb{P}\left\{\widehat{X}_O(t-\delta)>q-\epsilon\right\}\mathbb{P}\left\{Q_b(\infty)=d\right\}\\ &=\mathbb{P}\left\{\widehat{X}_O(t)>q-\epsilon\right\}\mathbb{P}\left\{Q_b(\infty)=d\right\}.\end{aligned}$$

The equalities above follows from the convergence of the sequence $\{\widehat{X}_O^\lambda\}$ to a continuous limit $\widehat{X}_O$ (see Theorem 3.1 in Whitt [2005]) and the continuity of the projection mapping in continuous limits (see e.g. §14 of Billingsley [1968]). In fact, the limit process in Whitt [2005] is a diffusion process with continuous drift and constant diffusion coefficients. In turn, for each $t>0$, the random variable $\widehat{X}_O(t)$ has a density (see, e.g., pages 368-369 in Karatzas and Shreve [1991]) so that

$$\begin{aligned}\liminf_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q, D_I^\lambda(t)=d\right\} &\geq \lim_{\epsilon\to 0}\mathbb{P}\left\{\widehat{X}_O(t)>q-\epsilon\right\}\mathbb{P}\left\{Q_b(\infty)=d\right\}\\ &=\mathbb{P}\left\{\widehat{X}_O(t)>q\right\}\mathbb{P}\left\{Q_b(\infty)=d\right\}.\end{aligned}\tag{EC.27}$$

To prove the other direction, note that on the set E_T, $\left\{\widehat{X}_O^\lambda(t)>q-\epsilon\right\}\subset\left\{\widehat{X}_O^\lambda(t-\delta)>q\right\}$. Arguing as before we have

$$\begin{aligned}\limsup_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q-\epsilon, D_I^\lambda(t)=d\right\} &\leq \lim_{\delta\to 0}\limsup_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t-\delta)>q\right\}\mathbb{P}\left\{D_I^\lambda(t)=d\right\}\\ &=\mathbb{P}\left\{\widehat{X}_O(t)>q\right\}\mathbb{P}\left\{Q_b(\infty)=d\right\},\end{aligned}$$

So that, upon taking limits with $\epsilon\downarrow 0$, we get from the $\mathcal{C}$-tightness of $\widehat{X}_O^\lambda$ that

$$\limsup_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q, D_I^\lambda(t)=d\right\}\leq\mathbb{P}\left\{\widehat{X}_O(t)>q\right\}\mathbb{P}\left\{Q_b(\infty)=d\right\}.\tag{EC.28}$$

It follows from (EC.27)-(EC.28) and (EC.5) that for all $q>0$ and $d\in\mathbb{Z}_+$,

$$\lim_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q, D_I^\lambda(t)=d\right\}=\mathbb{P}\left\{\widehat{X}_O(t)>q\right\}\mathbb{P}\left\{Q_b(\infty)=d\right\}$$
$$=\lim_{\lambda\to\infty}\mathbb{P}\left\{\widehat{X}_O^\lambda(t)>q\right\}\mathbb{P}\left\{D_I^\lambda(t)=d\right\}.$$

The independence with $\widehat{Q}_O^\lambda(t)$ replacing $\widehat{X}_O^\lambda(t)$ follows from the above argument noting that $\widehat{Q}_O^\lambda(t)=[\widehat{X}_O^\lambda(t)-(N_O^\lambda-R_O^\lambda)/\sqrt{\lambda}]^+$ where, by Assumption 1, $N_O^\lambda-R_O^\lambda=\varsigma\sqrt{(1-\nu)/\mu_O}\sqrt{\lambda}+o(\sqrt{\lambda})$. This concludes the proof. ■

Proof of Corollary 4.4: We first consider the virtual wait in stations I and O. To that end let $L_I^\lambda(t)$ and $L_O^\lambda(t)$ be the number of abandonments from station I and station O, respectively, by time t. Let $S_I^\lambda(t)$ and $S_O^\lambda(t)$ be the number of service completions at stations I and O, respectively by time t. Recall that $A^\lambda(t)$ is the number of exogenous arrivals to station I by time t and $A_O^\lambda(t)$ is the number of overflows by time t that correspond, in turn, to arrivals to station O. Following Talreja and Whitt [2009], given $t\geq 0$ and $u\geq 0$ we define $S_I^{\lambda,t}(t+u)$ and $L_I^{\lambda,t}(t+u)$ to be the service completion and abandonment by time $t+u$ assuming that the arrival process A^λ is stopped at time t. Similarly we define $S_O^{\lambda,t}(t+u)$ and $L_O^{\lambda,t}(t+u)$. Then, the virtual waiting time at a fixed time t satisfies

$$W_I^\lambda(t)=\inf\left\{u\geq 0: S_I^{\lambda,t}(t+u)-S_I^{\lambda,t}(t)+L_I^{\lambda,t}(t+u)-L_I^{\lambda,t}(t)\geq Q_I^\lambda(t)\right\}$$
$$W_O^\lambda(t)=\inf\left\{u\geq 0: S_O^{\lambda,t}(t+u)-S_O^{\lambda,t}(t)+L_O^{\lambda,t}(t+u)-L_O^{\lambda,t}(t)\geq Q_O^\lambda(t)\right\},$$

Due to the exponential service times and patience, $W_I^\lambda(t)$ depends on the past only via $X_I^\lambda(t)$ and $Q_I^\lambda(t)$. Similarly, $W_O^\lambda(t)$ depends on the past only via $X_O^\lambda(t)$ and $Q_O^\lambda(t)$. Since $Q_O^\lambda(t)=[X_O^\lambda(t)-N_O^\lambda]^+$ and $Q_I^\lambda(t)=[X_I^\lambda(t)-N_I^\lambda]^+$, the dependence is in fact only via $X_O^\lambda(t)$. Theorem 4.3 shows that $\widehat{X}_O^\lambda(t)$ is asymptotically independent of $D_I^\lambda(t)$ so the same is true for $\widehat{W}_O^\lambda(t)$. Consequently, we have for every bounded continuous function f that

$$\mathbb{E}\left[f(\widehat{W}^\lambda(t))\middle|D_I^\lambda(t)=0\right]=\mathbb{E}\left[f(\widehat{W}_O^\lambda(t))\right]\mathbb{P}\left\{D_I^\lambda(t)=0\right\}+o(1). \tag{EC.29}$$

By Theorem 4.1, and under the condition of the corollary, we have that $\widehat{Q}_I^\lambda\Rightarrow\bar{K}$ in $\mathcal{D}[0,\infty)$, where $\bar{K}$ is defined in the statement of the corollary. Also, by Theorem 4.1 it holds that $A_I^\lambda/\lambda=A^\lambda/\lambda-A_O^\lambda/\lambda\Rightarrow\nu e$

in $\mathcal{D}[0,\infty)$ and it is easy to show that, under the condition of the corollary, $S_I^\lambda/\lambda + L_I^\lambda/\lambda \Rightarrow \nu e$. Applying Theorem 3.1 in Talreja and Whitt [2009][3] we have that

$$\widehat{W}_I^\lambda \Rightarrow \widehat{W}_I = \bar{K}/\nu \quad \text{in } \mathcal{D}[0,\infty). \tag{EC.30}$$

Consequently,

$$\mathbb{E}\left[f(\widehat{W}_I^\lambda(t))\mathbb{1}\{D_I^\lambda(t)>0\}\right] = \mathbb{E}[f(\widehat{W}_I^\lambda(t))]\mathbb{P}\left\{D_I^\lambda(t)>0\right\} + o(1). \tag{EC.31}$$

Combining (EC.29) and (EC.31) we have that

$$\begin{aligned}\mathbb{E}\left[f(\widehat{W}^\lambda(t))\right] &= \mathbb{E}\left[f(\widehat{W}^\lambda(t))\middle|D_I^\lambda(t)>0\}\right]\mathbb{P}\left\{D_I^\lambda(t)>0\right\}\\ &\quad+\mathbb{E}\left[f(\widehat{W}^\lambda(t))\middle|D_I^\lambda(t)=0\right]\mathbb{P}\left\{D_I^\lambda(t)=0\right\}\\ &= \mathbb{E}\left[f(\widehat{W}_I^\lambda(t))\right]\mathbb{P}\left\{D_I^\lambda(t)>0\right\} + \mathbb{E}\left[f(\widehat{W}_O^\lambda(t))\right]\mathbb{P}\left\{D_I^\lambda(t)=0\right\} + o(1).\end{aligned} \tag{EC.32}$$

Noting that, by Theorem EC.1.4 and Corollary 4.2 we have $\mathbb{P}\left\{D_I^\lambda(t)=0\right\} = p_b^\lambda + o(1)$ for all $t>0$, we conclude that

$$\mathbb{E}\left[f(\widehat{W}^\lambda(t))\right] = \mathbb{E}\left[f(\widehat{W}_I^\lambda(t))\right](1-p_b^\lambda) + \mathbb{E}\left[f(\widehat{W}_O^\lambda(t))\right]p_b^\lambda + o(1).$$

We turn to prove the second part of the corollary. To that end, write

$$\mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A^\lambda(t)} f(\sqrt{\lambda}w_k^\lambda)\right] = \mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A_I^\lambda(t)} f(\sqrt{\lambda}w_{k,I}^\lambda)\right] + \mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda}w_{k,O}^\lambda)\right] \tag{EC.33}$$

Considering first the second element on the right hand side, note that

$$\mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda}w_{k,O}^\lambda)\right] = \mathbb{E}\left[\frac{A_O^\lambda(t)}{A^\lambda(t)}\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda}w_{k,O}^\lambda)\right].$$

By Theorem 4.1 and the strong law for renewal processes we have that, for each $t>0$, $\left|A_O^\lambda(t)/A^\lambda(t) - p_b^\lambda\right| \Rightarrow 0$, and the convergence also holds in expectation since $A_O^\lambda(t)/A^\lambda(t) \le 1$ for each $t>0$. Using the fact that f is bounded we have that

$$\mathbb{E}\left[\left|\frac{A_O^\lambda(t)}{A^\lambda(t)} - p_b^\lambda\right|\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda}w_{k,O}^\lambda)\right] = o(1)$$

[3] We note that Theorem 3.1 in Talreja and Whitt [2009] is not stated for queues with finite waiting room and, moreover, it requires that both S_I^λ/λ and L_I^λ/λ converge and not just the sum. Nevertheless, it is easy to verify that the result does apply here.

and, in turn, that

$$\mathbb{E}\left[\frac{A_O^\lambda(t)}{A^\lambda(t)}\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda}w_{k,O}^\lambda)\right] = p_b^\lambda \mathbb{E}\left[\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda}w_{k,O}^\lambda)\right] + o(1).$$

A similar analysis applies then to the first element on the right hand side of (EC.33). ■

Proof of Corollary 4.5: Note that by the functional strong law of large numbers applied to the Poisson process A^λ we have that $A^\lambda/\lambda \Rightarrow e$ in $\mathcal{D}[0,\infty)$. By Theorem 4.1 we have that $A_O^\lambda/\lambda \Rightarrow (1-\nu)e$ in $\mathcal{D}[0,\infty)$. In turn, we also have that $A_I^\lambda/\lambda \Rightarrow \nu e$. Finally, by Theorem 5.2 of Talreja and Whitt [2009] we also have that $\widehat{W}_O^\lambda \Rightarrow \widehat{W}_O = [\widehat{X}_O]^+/(1-\nu)$ where $\widehat{X}_O$ is the limit for the scaled and centered head-count process in the associated sequence of $GI/M/N_O^\lambda + M$ queues as in Theorem 3.1 of Whitt [2005]. In fact, Theorem 5.2 in Talreja and Whitt [2009] is for the $M/M/N+M$ queue but the result applies identically for the $GI/M/N+M$ replacing the appropriate limits for the head-count process of the $M/M/N+M$ queue (Theorem 5.1 there) with those of the $GI/M/N+M$ in Whitt [2005].

Combining all of the above and since all but $\widehat{W}_O^\lambda$ converge to non-random limits we can conclude (see Theorem 11.4.5 in Whitt [2002a]) the joint convergence

$$\left(\frac{A_I^\lambda}{\lambda}, \frac{A_O^\lambda}{\lambda}, \widehat{W}_I^\lambda, \widehat{W}_O^\lambda\right) \Rightarrow \left(\nu e, (1-\nu)e, \bar{K}/\nu, \widehat{W}_O\right) \quad \text{in } \mathcal{D}^4[0,\infty) \text{ as } \lambda \to \infty. \tag{EC.34}$$

From here the proof of the corollary follows from Corollary 4.4 by applying results in stochastic integration. Note first that we can write

$$\frac{1}{A_I^\lambda(t)}\sum_{k=1}^{A_I^\lambda(t)} f(\sqrt{\lambda}w_{k,I}^\lambda) = \frac{1}{A_I^\lambda(t)}\int_0^t f(\widehat{W}_I^\lambda(s-))dA_I^\lambda(s),$$

and

$$\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda}w_{k,O}^\lambda) = \frac{1}{A_O^\lambda(t)}\int_0^t f(\widehat{W}_O^\lambda(s-))dA_O^\lambda(s).$$

We will next use the following general result (whose proof appears in §EC.1.3)

LEMMA EC.1.10. Fix $t>0$. Let $\{(\Psi^\lambda, Y^\lambda)\}$ be a sequence of processes in $\mathcal{D}^2[0,t]$ such that, for all λ, Y^λ is increasing with $Y^\lambda(0)=0$ and such that $(\Psi^\lambda, Y^\lambda) \Rightarrow (\Psi, Y)$ as $\lambda \to \infty$ in the J_1 metric where Y is a continuous process. Let $f:\mathbb{R}\to\mathbb{R}_+$ be such that

$$\int_0^t \mathbb{1}\{\Psi(s-) \in disc\{f\}\}dY(s) = 0 \quad \text{almost surely,} \tag{EC.35}$$

where $disc\{f\}$ is the set of discontinuity points of the function f. Then for any $s \in [0,t]$,

$$\int_0^s f(\Psi^\lambda(u-))dY^\lambda(u) \Rightarrow \int_0^s f(\Psi(u-))dY(u), \text{ as } \lambda \to \infty.$$

Proceeding with the proof of Corollary 4.5 and fixing t, define for $0 \le s \le t$, the processes $Y_I^\lambda(s) := A_I^\lambda(s)/A_I^\lambda(t)$ and $Y_O^\lambda(s) = A_O^\lambda(s)/A_O^\lambda(t)$. By the strong law for renewal process we then have that, uniformly over compact subsets of $[0,t]$, both $Y_I^\lambda \Rightarrow e(\cdot)/t$ and $Y_O^\lambda \Rightarrow e(\cdot)/t$. Also, by the first part of the corollary, we have that $\widehat{W}_I^\lambda \Rightarrow \widehat{W}_I$ and $\widehat{W}_O^\lambda \Rightarrow \widehat{W}_O$. Since the limit processes are continuous this implies convergence together in $D^2[0,T]$ and, in particular, that each of the sequences $\{(\widehat{W}_I^\lambda, Y_I^\lambda)\}$ and $\{(\widehat{W}_O^\lambda, Y_O^\lambda)\}$ satisfies the conditions of Lemma EC.1.10. Moreover, since the function f in the corollary is assumed to be continuous (EC.35) holds trivially. Consequently, we conclude that

$$\frac{1}{A_I^\lambda(t)} \int_0^t f(\widehat{W}_I^\lambda(s-))dA_I^\lambda(s) \Rightarrow \frac{1}{t}\int_0^t f(\widehat{W}_I(s-))ds = \frac{1}{t}\int_0^t f(\widehat{W}_I(s))ds,$$

where the equalities hold almost surely using the boundedness of f, and similarly

$$\frac{1}{A_O^\lambda(t)} \int_0^t f(\widehat{W}_O^\lambda(s-))dA_O^\lambda(s) \Rightarrow \frac{1}{t}\int_0^t f(\widehat{W}_O(s-))ds = \frac{1}{t}\int_0^t f(\widehat{W}_O(s))ds.$$

Since the function f is bounded and since $A_I^\lambda(s)/A_I^\lambda(t) \le 1$ and $A_O^\lambda(s)/A_O^\lambda(t) \le 1$ for all $s \le t$, the convergence also holds in expectations. The proof of the corollary is thus complete. ■

EC.1.3. Proofs of auxiliary results

Proof of Lemma EC.1.2: Given the ordering in Lemma EC.1.1, the claim of the lemma is implied by the following result for $M/M/1$ queues: Fix $\epsilon > 0$ such that (EC.1) and (EC.2) hold. Then

$$\lim_{a\to\infty} \limsup_{\lambda\to\infty} \mathbb{P}\left\{ \sup_{0\le s\le T} Q^+_{\epsilon, Y_d^\lambda}(\lambda s) \ge a\sqrt{\lambda} \right\} = 0 \quad \text{and} \tag{EC.36}$$

$$\lim_{a\to\infty} \limsup_{\lambda\to\infty} \mathbb{P}\left\{ \sup_{\eta\le s\le T} Q^+_{\epsilon, Y_d^\lambda}(\lambda s) \ge a\log\lambda \right\} = 0, \quad \text{for all } \eta > 0. \tag{EC.37}$$

To establish (EC.37) it suffices to show that for any given $\xi > 0$, there exist $b(\xi)$ such that

$$\lim_{a\to\infty} \limsup_{\lambda\to\infty} \mathbb{P}\left\{ \sup_{b(\xi)/\sqrt{\lambda}\le s\le T} Q^+_{\epsilon, Y_d^\lambda}(\lambda s) > a\log\lambda \right\} \le \xi. \tag{EC.38}$$

At the end of the proof we will show how (EC.36) also follows from (EC.38).

Now, for $m \in \mathbb{R}_+$, let

$$\tau_m^\lambda(Y_d^\lambda) := \inf\left\{t \geq 0 : Q^+_{\epsilon,Y_d^\lambda}(\lambda t) \leq m\right\} \wedge T. \tag{EC.39}$$

We have that

$$\mathbb{P}\left\{\sup_{b(\xi)/\sqrt{\lambda}\leq s\leq T} Q^+_{\epsilon,Y_d^\lambda}(\lambda s) > a\log\lambda\right\} \leq \mathbb{P}\left\{\sup_{\tau_m^\lambda(Y_d^\lambda)\leq s\leq T} Q^+_{\epsilon,Y_d^\lambda}(\lambda s) > a\log\lambda\right\} + \mathbb{P}\left\{\tau_m^\lambda(Y_d^\lambda) > \frac{b(\xi)}{\sqrt{\lambda}}\right\}. \tag{EC.40}$$

We now treat each of the elements on the Right-Hand Side (RHS) of (EC.40), starting with the first element. We claim that

$$\mathbb{P}\left\{\sup_{\tau_m^\lambda(Y_d^\lambda)\leq s\leq T} Q^+_{\epsilon,Y_d^\lambda}(\lambda s) > a\log\lambda\right\} \leq \mathbb{P}\left\{\sup_{0\leq s\leq T} Q^+_{\epsilon,m}(\lambda s) > a\log\lambda\right\}. \tag{EC.41}$$

Indeed, if $\tau_m^\lambda(Y_d^\lambda) > T$, the set of times $\left\{s : \tau_m^\lambda(Y_d^\lambda) \leq s \leq T\right\}$ is empty so that the inequality holds trivially. For the other case, letting $\tilde{Y}_d^\lambda := Q^+_{\epsilon,Y_d^\lambda}(\lambda\tau_m^\lambda(Y_d^\lambda))$ (this is the state at the random time defined in (EC.39) and we set it to ∞ if $\tau_m^\lambda(Y_d^\lambda) > T$), we have that

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\tau_m^\lambda(Y_d^\lambda)\leq s\leq T} Q^+_{\epsilon,Y_d^\lambda}(\lambda s) > a\log\lambda;\ \tau_m^\lambda(Y_d^\lambda) \leq T\right\} \\
&\quad= \mathbb{P}\left\{\sup_{0\leq u\leq T-\tau_m^\lambda(Y_d^\lambda)} Q^+_{\epsilon,Y_d^\lambda}(\lambda(\tau_m^\lambda(Y_d^\lambda)+u)) > a\log\lambda;\ \tau_m^\lambda(Y_d^\lambda) \leq T\right\} \\
&\quad\leq \mathbb{P}\left\{\sup_{0\leq u\leq T} Q^+_{\epsilon,Y_d^\lambda}(\lambda(\tau_m^\lambda(Y_d^\lambda)+u)) > a\log\lambda;\ \tau_m^\lambda(Y_d^\lambda) \leq T\right\} \\
&\quad\overset{(a)}{=} \mathbb{P}\left\{\sup_{0\leq u\leq T} Q^+_{\epsilon,\tilde{Y}_d^\lambda}(\lambda u) > a\log\lambda;\ \tilde{Y}_d^\lambda \leq m\right\} \\
&\quad\overset{(b)}{=} \mathbb{P}\left\{\sup_{0\leq u\leq T} Q^+_{\epsilon,Y_d^\lambda\wedge m}(\lambda u) > a\log\lambda\right\}.
\end{aligned}$$

where Equality (a) uses the strong Markov property of the $M/M/1$ queue and Equality (b) follows from the fact that, on the event $\{\tilde{Y}_d^\lambda \leq m\}$, we have that $\tilde{Y}_d^\lambda = Y_d^\lambda \wedge m$. From the well-known monotonicity of the $M/M/1$ queue in its initial condition (see e.g. Chapter 9 of Ross [1996]), we then have

$$\mathbb{P}\left\{\sup_{\tau_m^\lambda(Y_d^\lambda)\leq s\leq T} Q^+_{\epsilon,Y_d^\lambda\wedge m}(\lambda s) > a\log\lambda\right\} \leq \mathbb{P}\left\{\sup_{0\leq s\leq T} Q^+_{\epsilon,m}(\lambda s) > a\log\lambda\right\},$$

which establishes (EC.41). To bound the right hand side of (EC.41), we have by Chapter VI.4 of Asmussen [2003], in particular by Problem 4.2 and Example 4.3 there, that

$$\lim_{a\to\infty}\limsup_{\lambda\to\infty}\mathbb{P}\left\{\sup_{0\le s\le T}Q^+_{\epsilon,m}(\lambda s)>a\log\lambda\right\}=0. \tag{EC.42}$$

We note that in Asmussen [2003] the cycles of the $M/M/1$ queue are started in 0, but the same applies if one considers a cycle as the time between consecutive visit to state m. We conclude from (EC.41) that

$$\lim_{a\to\infty}\limsup_{\lambda\to\infty}\mathbb{P}\left\{\sup_{\tau^\lambda_m(Y^\lambda_d)\le s\le T}Q^+_{\epsilon,Y^\lambda_d}(\lambda s)>a\log\lambda\right\}=0. \tag{EC.43}$$

This covers the first element on the right hand side of (EC.40), and we turn to the second element. To that end, fix $\xi>0$ and note that

$$\mathbb{P}\left\{\tau^\lambda_m(Y^\lambda_d)>\frac{b(\xi)}{\sqrt{\lambda}}\right\}\le\mathbb{P}\left\{\tau^\lambda_m(Y^\lambda_d)>\frac{b(\xi)}{\sqrt{\lambda}};Y^\lambda_d\le c\sqrt{\lambda}\right\}+\mathbb{P}\left\{Y^\lambda_d>c\sqrt{\lambda}\right\}. \tag{EC.44}$$

By the assumptions of the lemma, given ξ we can choose $c(\xi)$ so that $\mathbb{P}\left\{Y^\lambda_d>c(\xi)\sqrt{\lambda}\right\}\le\xi$. By the monotonicity of the $M/M/1$ queue in the initial condition we have that, on the event $\left\{Y^\lambda_d\le c(\xi)\sqrt{\lambda}\right\}$, $\tau^\lambda_m(Y^\lambda_d)\le_{st}\tau^\lambda_m(c(\xi)\sqrt{\lambda})$. Namely,

$$\mathbb{P}\left\{\tau^\lambda_m(Y^\lambda_d)>\frac{b(\xi)}{\sqrt{\lambda}},Y^\lambda_d\le c(\xi)\sqrt{\lambda}\right\}\le\mathbb{P}\left\{\tau^\lambda_m(c(\xi)\sqrt{\lambda})>\frac{b(\xi)}{\sqrt{\lambda}}\right\}.$$

recall that Q^+_ϵ is an $M/M/1$ queue with arrival rate $\nu+\epsilon<1$ and service rate 1. By basic results for the $M/M/1$ queue, see e.g., Proposition 3.3.1 in Meyn [2008] (note that $T^*(x)$ in that proposition is defined in (3.5), page 63 of that reference),

$$\mathbb{E}\left[\tau^\lambda_m(c(\xi)\sqrt{\lambda})\right]\le\frac{1}{\lambda}\frac{c(\xi)\sqrt{\lambda}}{1-\nu-\epsilon}=\frac{c(\xi)}{\sqrt{\lambda}(1-\nu-\epsilon)},$$

where the division by λ in the inequality above is due to the time being scaled by a factor λ. Then by Markov's inequality,

$$\mathbb{P}\left\{\tau^\lambda_m(c\sqrt{\lambda})>\frac{b(\xi)}{\sqrt{\lambda}}\right\}\le\frac{c(\xi)}{b(\xi)(1-\nu-\epsilon)}.$$

We can now choose $b(\xi)$ large enough so that $\mathbb{P}\left\{\tau^\lambda_m(Y^\lambda_d)>b(\xi)/\sqrt{\lambda},Y^\lambda_d\le c(\xi)\sqrt{\lambda}\right\}\le\xi$ for all λ large enough. Plugging this into (EC.44) we then have that

$$\limsup_{\lambda\to\infty}\mathbb{P}\left\{\tau^\lambda_m(Y^\lambda_d)>\frac{b(\xi)}{\sqrt{\lambda}}\right\}\le 2\xi. \tag{EC.45}$$

Since for each $\xi > 0$ we can find such $b(\xi)$, we can plug (EC.45) together with (EC.43) into (EC.40) to conclude the proof of (EC.38). We turn now to prove (EC.36).

To that end, we note that

$$\mathbb{P}\left\{\sup_{0\le s\le T} Q^+_{\epsilon,Y_d^\lambda}(\lambda s)\ge M\sqrt{\lambda}\right\}\le \mathbb{P}\left\{\sup_{b(\xi)/\sqrt{\lambda}\le s\le T} Q^+_{\epsilon,Y_d^\lambda}(\lambda s)\ge M/2\sqrt{\lambda}\right\} + \mathbb{P}\left\{\sup_{0\le s\le b(\xi)/\sqrt{\lambda}} Q^+_{\epsilon,Y_d^\lambda}(\lambda s)\ge M/2\sqrt{\lambda}\right\}.$$

The first element on the RHS was treated in (EC.38). For the second element we can use a crude bound. Note that for all $s\le b(\xi)/\sqrt{\lambda}$, $Q^+_{\epsilon,Y_d^\lambda}(\lambda s)\le_{st} Y_d^\lambda + \mathcal{N}(\lambda(\nu+\epsilon)b(\xi)/\sqrt{\lambda})$, where $\mathcal{N}(\cdot)$ is a unit rate Poisson process (i.e, the state at time s is smaller than the initial condition plus all the arrivals up to time $b(\xi)/\sqrt{\lambda}\ge s$). Since both $Y_d^\lambda = O_P(\sqrt{\lambda})$ and $\mathcal{N}(\lambda(\nu+\epsilon)b(\xi)/\sqrt{\lambda}) = O_P(\sqrt{\lambda})$ it follows that

$$\lim_{M\to\infty}\limsup_{\lambda\to\infty}\mathbb{P}\left\{\sup_{0\le s\le b(\xi)/\sqrt{\lambda}} Q^+_{\epsilon,Y_d^\lambda}(\lambda s)\ge M/2\sqrt{\lambda}\right\}=0.$$

■

Proof of Lemma EC.1.3: Let $\check{\tau}^\lambda_M(Y_d^\lambda) := \inf\left\{t\ge 0: D_I^\lambda(t)\ge M\sqrt{\lambda}\right\}\wedge T$. Recall that D_I^λ is a state-dependent BD process with state space $d\in\{0,1,\ldots,N_I^\lambda+K_I^\lambda\}$, having death rate λ and birth rate (in state d) $\mu(N_I^\lambda-(d-K_I^\lambda)^+)+\theta(K_I^\lambda-d)^+$ so that for all $d\le M\sqrt{\lambda}+1$

$$\mu(N_I^\lambda-(d-K_I^\lambda)^+)+\theta(K_I^\lambda-d)^+\ge\lambda(\nu-\epsilon).$$

The process $D_I^\lambda(t)$ can be constructed on the same sample space with the $M/M/1$ queue, $Q^-_{\epsilon,Y_d^\lambda}$ so that

$$D_I^\lambda(t)\ge Q^-_{\epsilon,Y_d^\lambda}(\lambda t),\ \text{for all } t\le\check{\tau}^\lambda_M(Y_d^\lambda).$$

The comparison does not necessarily hold after $\check{\tau}^\lambda_M(Y_d^\lambda)$. The simple coupling argument is omitted and we refer the reader to the proof of Lemma EC.1.11 where a very similar argument is used. Note now that for $t\le T$,

$$\begin{aligned}\mathbb{P}\left\{D_I^\lambda(t)\ge d_1\right\} &= \mathbb{P}\left\{D_I^\lambda(t)\ge d_1, T\le\check{\tau}^\lambda_M(Y_d^\lambda)\right\}+\mathbb{P}\left\{D_I^\lambda(t)\ge d_1, T>\check{\tau}^\lambda_M(Y_d^\lambda)\right\}\\ &\ge \mathbb{P}\left\{Q^-_{\epsilon,Y_d^\lambda}(\lambda t)\ge d_1, T\le\check{\tau}^\lambda_M(Y_d^\lambda)\right\}\end{aligned}$$

$$\geq \mathbb{P}\left\{Q^{-}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1\right\} - \mathbb{P}\left\{\check{\tau}^{\lambda}_M(Y^{\lambda}_d) < T\right\}.$$

To conclude the proof we only need to show that for all $t \leq T$,

$$\lim_{M\to\infty} \limsup_{\lambda\to\infty} \mathbb{P}\left\{\check{\tau}^{\lambda}_M(Y^{\lambda}_d) < T\right\} = 0. \tag{EC.46}$$

This, however, follows by noting that

$$\begin{aligned}\lim_{M\to\infty} \limsup_{\lambda\to\infty} \mathbb{P}\left\{\check{\tau}^{\lambda}_M(Y^{\lambda}_d) < T\right\} &\leq \lim_{M\to\infty} \limsup_{\lambda\to\infty} \mathbb{P}\left\{\sup_{0\leq s\leq T} D^{\lambda}_I(s) \geq M\sqrt{\lambda}\right\} \\ &\leq \lim_{M\to\infty} \limsup_{\lambda\to\infty} \mathbb{P}\left\{\sup_{0\leq s\leq T} Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda s) \geq M\sqrt{\lambda}\right\} = 0,\end{aligned}$$

where the second inequality follows from Lemma EC.1.1 and the last equality follows from (EC.37). ■

Proof of Lemma EC.1.5: We prove the result only of Q^{+}_{ϵ}. The proof is identical for Q^{-}_{ϵ}. To that end, let $\tau^{\lambda}_m(Y^{\lambda}_d)$ be defined as in (EC.39) and note that

$$\mathbb{P}\left\{Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1\right\} = \mathbb{P}\left\{Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1, \tau^{\lambda}_m(Y^{\lambda}_d) \leq b\right\} + \mathbb{P}\left\{Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1, \tau^{\lambda}_m(Y^{\lambda}_d) > b\right\}.$$

Since, for any $b > 0$, $\limsup_{\lambda\to\infty} \mathbb{P}\left\{\tau_m(Y^{\lambda}_d) > b\right\} = 0$ (see equation (EC.45) and the argument leading to it), we have that

$$\mathbb{P}\left\{Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1\right\} = \mathbb{P}\left\{Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1, \tau^{\lambda}_m(Y^{\lambda}_d) \leq b\right\} + o(1),$$

so that it suffices to consider $\mathbb{P}\left\{Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1, \tau^{\lambda}_m(Y^{\lambda}_d) \leq b\right\}$. By the strong Markov property of the $M/M/1$ queue we can write

$$\begin{aligned}\mathbb{P}\left\{Q^{+}_{\epsilon,Y^{\lambda}_d}(\lambda t) \geq d_1, \tau^{\lambda}_m(Y^{\lambda}_d) \leq b\right\} &= \mathbb{P}\left\{Q^{+}_{\epsilon,\tilde{Y}^{\lambda}_d}(\lambda(t-\tau^{\lambda}_m(Y^{\lambda}_d))) \geq d_1, \tau^{\lambda}_m(Y^{\lambda}_d) \leq b\right\} \\ &= \mathbb{P}\left\{Q^{+}_{\epsilon,\tilde{Y}^{\lambda}_d}(\lambda(t-\tau^{\lambda}_m(Y^{\lambda}_d))) \geq d_1\right\} + o(1),\end{aligned}$$

where $\tilde{Y}^{\lambda}_d$ is the state at the random time $\tau^{\lambda}_m(Y^{\lambda}_d)$, i.e, $\tilde{Y}^{\lambda}_d = Q^{+}_{\epsilon,Y^{\lambda}_d}(\tau^{\lambda}_m(Y^{\lambda}_d))$. The last equality follows again from the fact that $\mathbb{P}\left\{\tau^{\lambda}_m(Y^{\lambda}_d) > b\right\} \to 0$. Now,

$$\begin{aligned}\mathbb{P}\left\{Q^{+}_{\epsilon,\tilde{Y}^{\lambda}_d}(\lambda(t-\tau^{\lambda}_m(Y^{\lambda}_d))) \geq d_1\right\} &= \mathbb{P}\left\{Q^{+}_{\epsilon,\tilde{Y}^{\lambda}_d}(\lambda(t-\tau^{\lambda}_m(Y^{\lambda}_d))) \geq d_1, \tilde{Y}^{\lambda}_d \leq 2m\right\} \\ &\quad + \mathbb{P}\left\{Q^{+}_{\epsilon,\tilde{Y}^{\lambda}_d}(\lambda(t-\tau^{\lambda}_m(Y^{\lambda}_d))) \geq d_1, \tilde{Y}^{\lambda}_d > 2m\right\}.\end{aligned}$$

Since $\mathbb{P}\left\{\tau_m^\lambda(Y_d^\lambda) > b\right\} \to 0$ we also have that $\mathbb{P}\left\{\tilde{Y}_d^\lambda > 2m\right\} \to 0$. In turn, it suffices to consider the first element on the RHS above. Since $\tau_m^\lambda(Y_d^\lambda) \leq T$ by definition and since we consider the event on which the initial condition $\tilde{Y}_d^\lambda \leq 2m$ (which does not change with λ) we have that as $\lambda \to \infty$,

$$Q^+_{\epsilon,\tilde{Y}_d^\lambda}(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \Rightarrow Q_\epsilon^+(\infty),$$

and in turn that

$$\mathbb{P}\left\{Q^+_{\epsilon,\tilde{Y}_d^\lambda}(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \geq d_1, \tilde{Y}_d^\lambda \leq 2m\right\} \to \mathbb{P}\left\{Q_\epsilon^+(\infty) \geq d_1\right\}.$$

This concludes the proof. ■

Proof of Lemma EC.1.8 We start with the proof of (EC.14). The first part is straightforward. Indeed, since ξ_k^λ is exponential with rate $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$ we have that

$$Var(\xi_k^\lambda/\sqrt{\lambda}) = \frac{1}{\lambda}\frac{\lambda^2}{(\mu_I N_I^\lambda + \theta K_I^\lambda)^2},$$

so that, for each $t > 0$,

$$\lim_{\lambda\to\infty}\sum_{k=1}^{\lceil \lambda t\rceil} Var(\xi_k^\lambda/\sqrt{\lambda}) = \lim_{\lambda\to\infty}\frac{\lceil \lambda t\rceil}{\lambda}\frac{\lambda^2}{(\mu_I N_I^\lambda + \theta K_I^\lambda)^2} = \frac{1}{\nu^2}t, \tag{EC.47}$$

where we used the fact that, by our assumptions $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda \to \nu$. Note that, since both the pre-limit functions above are monotone increasing and since the limit is continuous, the pointwise convergence implies uniform convergence. For the second part of (EC.14) we have the following auxiliary result whose proof appears at the end of this section.

LEMMA EC.1.11. As $\lambda \to \infty$,

$$\sqrt{\lambda}\left(\mathbb{E}\left[\bar{T}_1^\lambda\right] - \left(1 - \frac{\mu_I N_I^\lambda + \theta K_I^\lambda}{\lambda}\right)^{-1}\right) \to 0, \tag{EC.48}$$

$$\sqrt{\lambda}\left((1 - \rho_d^\lambda) - \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\lambda}\right) \to 0, \tag{EC.49}$$

and

$$Var[\bar{T}_1^\lambda] \to \frac{1+\nu}{(1-\nu)^3}, \tag{EC.50}$$

and consequently

$$\sum_{k=1}^{\lceil \lambda \cdot \rceil} Var\left(\frac{\bar{T}_k^\lambda}{\sqrt{\lambda}}\right) \Rightarrow \frac{1+\nu}{(1-\nu)^3} e, \tag{EC.51}$$

in $\mathcal{D}[0,\infty)$.

Using (EC.47) and (EC.51) we have immediately that

$$\sum_{k=1}^{\lceil \lambda \cdot \rceil} Var\left(\frac{T_k^\lambda}{\sqrt{\lambda}}\right) \Rightarrow \left(\frac{1+\nu}{(1-\nu)^3} + \frac{1}{\nu^2}\right) e, \tag{EC.52}$$

in $\mathcal{D}[0,\infty)$, which proves (EC.14). We can now apply the FCLT for double arrays (see e.g. Theorem 2.3.9 in the internet supplement to Whitt [2002a]) to conclude that, as $\lambda \to \infty$,

$$\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lceil \lambda \cdot \rceil} \left(T_k^\lambda - \mathbb{E}\left[T_k^\lambda\right]\right) \Rightarrow \check{\sigma} B \quad \text{as} \quad \lambda \to \infty,$$

in $\mathcal{D}[0,\infty)$ where $\check{\sigma}^2 = \frac{1+\nu}{(1-\nu)^3} + \frac{1}{\nu^2}$. By Theorem 1 of Iglehart and Whitt [1971] the convergence of the partial sums implies convergence of the corresponding renewal process so that

$$\frac{1}{\sqrt{\lambda}} \left(R^\lambda(\lambda \cdot) - \lambda / \mathbb{E}\left[T_k^\lambda\right] \cdot\right) \Rightarrow \sqrt{\nu(1-\nu)} \check{\sigma} B \quad \text{as} \quad \lambda \to \infty,$$

in $\mathcal{D}[0,\infty)$. From here it also follows directly that, uniformly on compact subsets,

$$R^\lambda(\lambda \cdot)/\lambda - \cdot / \mathbb{E}\left[T_k^\lambda\right] \Rightarrow 0 \quad \text{as} \quad \lambda \to \infty,$$

in $\mathcal{D}[0,\infty)$. Since $\mathbb{E}\left[T_k^\lambda\right] \to 1/\nu + 1/(1-\nu) = 1/(\nu(1-\nu))$, we have that $R(\lambda \cdot)/\lambda \Rightarrow \nu(1-\nu)e$ in $\mathcal{D}[0,\infty)$. The convergence in (EC.14) now follows from the marginal convergence of the components because all limits are non-random (see Theorem 11.4.5 in Whitt [2002a]). We now use (EC.47), (EC.51) and (EC.49) and apply the FCLT for double arrays to each of the two (independent) processes

$$\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} \left(\rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda}\right), \text{ and } \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} (1-\rho_d^\lambda) \left(\bar{T}_k^\lambda - \mathbb{E}\left[\bar{T}_k^\lambda\right]\right)$$

to obtain the joint convergence of

$$\left(\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} \left(\rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda}\right), \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} (1-\rho_d^\lambda) \left(\bar{T}_k^\lambda - \mathbb{E}\left[\bar{T}_k^\lambda\right]\right), \frac{R(\lambda \cdot)}{\lambda}\right) \Rightarrow \left(B^1, \sqrt{\frac{1+\nu}{1-\nu}} B^2, \nu(1-\nu) e\right),$$

in $\mathcal{D}[0,\infty)$ where B^1 and B^2 are standard independent Brownian motions. Equation (EC.15) now follows from the random time change theorem; see e.g. §17 of Billingsley [1968]. ■

Proof of Lemma EC.1.9 The first part follows from (EC.48) and (EC.49) after basic algebraic manipulations. Equation (EC.16) than follows since $R^\lambda(\lambda\cdot)/\lambda \Rightarrow \nu(1-\nu)e$ as proved within the proof of Lemma EC.1.8. ■

Proof of Lemma EC.1.10: Consider first the (sequence of) processes $M^\lambda(s),\ 0 \le s \le t$ defined by

$$M^\lambda(s) := \int_0^s \Psi^\lambda(u-)dY^\lambda(u),\ 0 \le s \le t,$$

and let $M(s) = \int_0^s \Psi(u-)dY(u)$. Then, from Theorem 0 in Jakubowski [1996] it follows, under certain conditions to be discussed shortly, that $M^\lambda \Rightarrow M$ in $\mathcal{D}[0,t]$. The conditions in that theorem refer to a certain UT property of the process Y^λ (see the discussion in Jakubowski [1996] immediately after the statement of Theorem 0 there). However, this condition holds trivially when (as in our case) Y^λ is, for each λ, an increasing process and such that $Y^\lambda \Rightarrow Y$, which implies, in particular, that the sequence Y^λ is stochastically bounded. In addition, the martingale condition in Theorem 0 in Jakubowski [1996] holds trivially, as one can always use the self filtration of the process $(\Psi^\lambda, Y^\lambda)$, for each λ. The process Ψ^λ is then adapted to this filtration and, being an increasing process, Y^λ is a semi-martingale with respect to this filtration. Thus, it indeed holds that $M^\lambda \to M$.

We next show that the convergence holds if one replaces the integrand $\Psi^\lambda(s-)$ with $f(\Psi^\lambda(s-))$, where f satisfies (EC.35). Note that, for all λ, $t > 0$ and Borel measurable set $\mathcal{B}$, we have that

$$\mathcal{M}^\lambda(\mathcal{B}) := \int_0^t \mathbb{1}\{\Psi^\lambda(s-) \in \mathcal{B}\}dY^\lambda(s)$$

is a (random measure) on $\mathbb{R}$; see e.g. **?**. By the first part of this proof it holds that

$$\int_0^t f(\Psi^\lambda(s-))dY^\lambda(s) \Rightarrow \int_0^t f(\Psi(s-))dY(s),$$

for every continuous function $f : \mathbb{R} \to \mathbb{R}_+$. Thus, the sequence of random measures $\mathcal{M}^\lambda(\cdot)$ converges weakly to $\mathcal{M}(\cdot)$, where

$$\mathcal{M}(\mathcal{B}) = \int_0^t \mathbb{1}\{X(s-) \in \mathcal{B}\}dY(s).$$

The condition (EC.35) implies that $\mathcal{M}(\cdot)$ assigns zero measure to the family of discontinuity points of f so that, from the generalized version of continuous mapping theorem (see, e.g., Theorem 3.4.3 in Whitt [2002b]), we deduce the convergence

$$\int_0^t f(\Psi^\lambda(s-))dY^\lambda(s) \Rightarrow \int_0^t f(\Psi(s-))dY(s),$$

for all functions f that satisfy condition (EC.35). ■

Proof of Lemma EC.1.11 Recall that $\bar{T}^\lambda$ has the distribution of the length of a busy period of the process $D_I^{s,\lambda}(\cdot)$, i.e, $\bar{T}^\lambda = T^\lambda - \xi^\lambda$ for T^λ that has the distribution of the busy cycle and ξ^λ that has the distribution of the idle period. To obtain bounds for the busy period we will couple the process $D_I^{s,\lambda}(\cdot)$ with two queues.

Specifically, let $\check{T}^\lambda$ have the distribution of a busy period in an $M/M/1$ queue with service rate 1 and arrival rate $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$. Let $\check{\rho}^\lambda$ be the utilization in this $M/M/1$ queue. Let $\tilde{T}^\lambda$ have the distribution of a busy period in an $M/M/1$ queue with service rate 1 and arrival rate $(\mu_I N_I^\lambda/\lambda + \theta K_I^\lambda - 2K\log(\lambda))/\lambda$ (note that for all λ large enough this is a positive number and it is also strictly smaller than 1). Let $\tilde{\rho}^\lambda$ be the utilization in this $M/M/1$ queue and let $\tilde{M}^\lambda$ be a random variable with the distribution of the busy-period maximum in such an $M/M/1$ queue.

We will next use a coupling argument to show that

$$\mathbb{E}\left[\tilde{T}^\lambda \mathbb{1}\left\{\tilde{M}^\lambda \leq K\log(\lambda)\right\}\right] \leq \mathbb{E}\left[\bar{T}^\lambda\right] \leq \mathbb{E}\left[\check{T}^\lambda\right] \tag{EC.53}$$

and similarly that

$$\mathbb{E}\left[(\tilde{T}^\lambda)^2 \mathbb{1}\left\{\tilde{M}^\lambda \leq K\log(\lambda)\right\}\right] \leq \mathbb{E}\left[(\bar{T}^\lambda)^2\right] \leq \mathbb{E}\left[(\check{T}^\lambda)^2\right]. \tag{EC.54}$$

Let $\check{Q}^\lambda(t)$ be the queue length of the $M/M/1$ queue with service rate 1 and arrival rate $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$ and let $\tilde{Q}^\lambda(t)$ be defined similarly with the arrival rate $(\mu_I N_I^\lambda/\lambda + \theta K_I^\lambda - 2K\log(\lambda))/\lambda$. Then, we start at time 0 all the queues with the server busy and zero customers in queue. Namely, set $\check{Q}^\lambda(0) = \tilde{Q}^\lambda(0) = D_I^{s,\lambda}(0) = 1$. Then, we will couple them until the first of them hits 0. We generate "events" using a Poisson process with rate $b^\lambda := 1 + (\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$. We determine whether an event is an arrival or a service completion using the same sequence of uniform $[0,1]$ random variables for all the three queues (for each

event we use a different uniform random variable and these uniforms are IID). The k^{th} event trigger an arrival in $D_I^{s,\lambda}$ when in state d if $U_k \in [0, ((\mu(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(d \wedge K_I^\lambda))/\lambda)/b^\lambda)$. It is a service completion if $U_k \in [1 - 1/b^\lambda, 1]$ and $D_I^{s,\lambda}$ stays in its state d otherwise. Similarly, an event triggers an arrival in $\check{Q}_1^\lambda$ if $U_k \in [0, ((\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda)/b^\lambda)$. It trigger a service completion if $U_k \in [1 - 1/b^\lambda, 1]$ and it stays put otherwise. Finally, an event triggers an arrival in $\tilde{Q}_1^\lambda$ if $U_k \in [0, ((\mu_I N_I^\lambda + \theta K_I^\lambda - 2K\log(\lambda))/\lambda)/b^\lambda)$. It triggers a service completion if $U_k \in [1 - 1/b^\lambda, 1]$ and it stays put otherwise.

Let $\tilde{\tau}^\lambda = \inf\left\{t \geq 0 : \tilde{Q}^\lambda(t) = 0 \text{ or } \tilde{Q}^\lambda(t) \geq K\log(\lambda)\right\}$. Then with the above sample-path construction is necessarily holds that:

(i) for all $t \geq 0$, $D_I^{s,\lambda}(t) \leq \check{Q}^\lambda(t)$ and

(ii) for all $t \leq \tilde{\tau}^\lambda$, $\tilde{Q}^\lambda(t) \leq D_I^{s,\lambda}(t) \leq \check{Q}^\lambda(t)$.

From here it also follows that, under this construction

$$\tilde{T}^\lambda 1\left\{\tilde{M}^\lambda \leq K\log\lambda\right\} \leq \bar{T}^\lambda \leq \check{T}^\lambda$$

almost surely. Taking squares and applying expectations on both sides (ny Proposition 8.10 in Asmussen [2003] these expectations are finite), we have (EC.53) and (EC.54).

We next use (EC.53) and (EC.54) to complete the proof of the lemma. By known results for the busy period of the $M/M/1$ queue (see e.g. Proposition 8.10 in Asmussen [2003]) we have

$$\mathbb{E}\left[\check{T}^\lambda\right] = \frac{1}{(1 - (\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda)} \text{ and } \mathbb{E}\left[\tilde{T}^\lambda\right] = \frac{1}{(1 - (\mu_I N_I^\lambda + \theta K_I^\lambda - 2K\log(\lambda))/\lambda)},$$

so that $\sqrt{\lambda}\left(\mathbb{E}\left[\check{T}^\lambda\right] - \mathbb{E}\left[\tilde{T}^\lambda\right]\right) \to 0$, and, to establish the first part of the lemma, it suffices to show that

$$\sqrt{\lambda}\mathbb{E}\left[\tilde{T}^\lambda \mathbb{1}\left\{\tilde{M}^\lambda > K\log(\lambda)\right\}\right] \to 0, \text{ as } \lambda \to \infty.$$

By Holder's inequality

$$\mathbb{E}\left[\tilde{T}^\lambda \mathbb{1}\left\{\tilde{M}^\lambda > K\log(\lambda)\right\}\right] \leq \sqrt{\mathbb{E}\left[(\tilde{T}^\lambda)^2\right]}\sqrt{\mathbb{P}\left\{\tilde{M}^\lambda > K\log(\lambda)\right\}}.$$

It is known (see again Asmussen [2003]) that $\mathbb{E}\left[(\tilde{T}^\lambda)^2\right] = (1 + \tilde{\rho}^\lambda)/(1 - \tilde{\rho}^\lambda)^3$ so that, since $\tilde{\rho}^\lambda \to \nu < 1$

by Assumption 1, we have $\limsup_{\lambda\to\infty}\mathbb{E}\left[(\tilde{T}^\lambda)^2\right]<\infty$. Note that $\tilde{M}^\lambda\leq_{st}M^\lambda:=\sup_{n\geq 0}S_n^\lambda$ where S_n^λ is random walk (starting at 1) that increases by one with probability

$$\phi^\lambda=\frac{(\mu_I N_I^\lambda+\theta K_I^\lambda-2K\log(\lambda)/\lambda)}{\lambda+(\mu_I N_I^\lambda+\theta K_I^\lambda-2K\log(\lambda)/\lambda)}$$

and decreases by one with probability $1-\phi^\lambda$. This is a Gambler's-ruin type problem and from basic arguments is follows that

$$\mathbb{P}\left\{M^\lambda>K\log(\lambda)+1\right\}=\left(\frac{\phi^\lambda}{1-\phi^\lambda}\right)^{2K\log\lambda};$$

see e.g. Chapter 7.3 of Ross [1996]. By Assumption 1, $\phi^\lambda\to 1/(1+\nu)<1$. In turn, there exists a constant $c<1$ such that $\mathbb{P}\left\{M^\lambda>K\log(\lambda)+1\right\}\leq c^{2K\log\lambda}$ for all λ large enough. Recalling that $\tilde{M}^\lambda\leq_{st}M^\lambda$ and choosing K large enough, we conclude that

$$\sqrt{\lambda}\sqrt{\mathbb{P}\left\{\tilde{M}^\lambda>K\log(\lambda)\right\}}\to 0.$$

Similar arguments are applied to the variance terms. The only difference is that when applying Holder's inequality we will use the fact that, since $\tilde{\rho}^\lambda\to\nu$, the sequence $\{\mathbb{E}[(\tilde{T}^\lambda)^4],\ \lambda\geq 0\}$ convergence so that, in particular, $\limsup_{\lambda\to\infty}\mathbb{E}\left[(\tilde{T}^\lambda)^4\right]<\infty$. The fact that the moments converge follows from the convergence of the characteristic function together with its derivatives at 0 of all orders as $\lambda\to\infty$ and $\tilde{\rho}^\lambda\to\nu$. The convergence of the characteristic function is easily verifiable using the explicit expressions; see Proposition 8.10 of Asmussen [2003].

To conclude the proof we show that (EC.49) follows from (EC.48). Indeed, by basic regenerative process arguments (applied to the regenerative process $D_I^{s,\lambda}$) we have that

$$\rho_d^\lambda=\frac{\mathbb{E}\left[\bar{T}^\lambda\right]}{\mathbb{E}\left[\bar{T}^\lambda\right]+\mathbb{E}\left[\xi^\lambda\right]}.$$

Recalling that $\mathbb{E}\left[\xi^\lambda\right]=\frac{1}{(\mu_I N_I^\lambda+\theta K_I^\lambda)/\lambda}$ and using (EC.48) the result by basic algebraic manipulations. ■

EC.2. The Multi-Class Setting

Thus far we considered a setting in which the outside provider uses a dedicated pool of servers for the overflow input. With this configuration, analyzing a system with multiple in-house call centers is identical

to analyzing multiple independent subsystems consisting of an in-house call center and its dedicated station at the outside provider's facility.

In this section we consider the case in which the outside provider serves multiple input streams in a common facility. Station O can then be thought of as a call center with multiple customer classes and possibly multiple agent pools. An example having a single agent pool is depicted in Figure 1(b). More specifically, each overflow stream can be thought of as a customer class. There might be different agent groups that differ by their skills set, and hence have different levels of flexibility. In such a multi-class multi-pool configuration, the outsourcer needs to determine the policy for real-time prioritization of customers.

For the system in Figure 1(b), the controller has to determine: (i) what to do with an arriving call if there is an available agent upon its arrival (admit it to service or reserve the capacity for other customer classes), and (ii) which customer to admit to service when an agent becomes available and there are customers waiting in multiple queues. We refer the reader to Gurvich and Whitt [2010] for a more detailed description of the underlying queueing network.

Fortunately, our results for the base model extend to this more complex setting. Below we highlight what are the corresponding mathematical statements and what are the assumptions that have to be imposed on the prioritization rule in station O. To formally state the results, we let $\mathcal{M} := \{1, \ldots, M\}$ be the set of in-house centers, i.e, the sources of overflows for station O. The process A_m is then the Poisson process of exogenous arrivals to (in-house) center $m \in \mathcal{M}$, with rate λ_m. We denote the aggregate arrival rate by $\lambda = \sum_m \lambda_m$. As before, we add the superscript λ to all quantities that change along the sequence of systems. We will assume that $\lambda_m/\lambda = a_m > 0$, i.e., the fractions a_m do not scale with λ.

The service rate at station m is given by μ_m and we let θ_m be the patience rate of the customers arriving to station m. Let $(N_m^\lambda, K_m^\lambda)$ be the capacity and threshold pair for station $m \in \mathcal{M}$. The resource pooling condition (item (1) in Assumption 1) is assumed to hold for each of the in-house centers, i.e,

$$\lim_{\lambda\to\infty} \frac{\mu_m N_m^\lambda + \theta K_m^\lambda}{\lambda_m} = \nu_m < 1, \ m \in \mathcal{M}. \tag{EC.55}$$

Let $A_{m,O}^\lambda$ be the overflow process from station m, and define the scaled process

$$\widehat{A}_{m,O}^\lambda(t) = \frac{A_{m,O}^\lambda(t) - (\lambda_m - \mu_m N_m^\lambda - \theta_m K_m^\lambda)t}{\sqrt{\lambda_m}}.$$

For state descriptors, we use $X_{m,I}(t)$ to denote the number of customers present in station $m \in \mathcal{M}$ at time t. In station O there is a queue for each overflow stream (for each customer class) and we let $X_{m,O}(t)$ and $Q_{m,O}(t)$ be, respectively, the total head count of "class-m" customers in station O at time t, and the number in queue there. As in Assumption 1, we will assume that station O uses a square-root safety staffing rule.

For in-house station $m \in \mathcal{M}$ we define the scaled and centered process $\widehat{X}^{\lambda}_{m,I} := (X^{\lambda}_{m,I} - (N^{\lambda}_{m} + K^{\lambda}_{m}))/\sqrt{\lambda}$. For station O and class k we denote by $\widehat{Q}^{\lambda}_{m,O}$ the scaled queue length process and by $\widehat{X}^{\lambda}_{m,O}$ the scaled and centered head-count process. The square-root staffing rule, as well as the centering of the head-count processes $X^{\lambda}_{m,O}$, need to be defined carefully in the multi-class multi-pool setting. Such details are not central for our results below and we refer the reader to §2.1 of Gurvich and Whitt [2009]. We will be assuming, analogously to (12), that the initial conditions converge, i.e, that

$$(\widehat{Q}^{\lambda}_{m,O}(0), \widehat{X}^{\lambda}_{m,O}(0), \widehat{X}^{\lambda}_{m}(0); m \in \mathcal{M}) \Rightarrow (\widehat{Q}_{m,O}(0), \widehat{X}_{m,O}(0), \widehat{X}_{m}(0); m \in \mathcal{M}) \text{ as } \lambda \to \infty. \tag{EC.56}$$

Since the in-house call centers are independent of each other, Theorem 4.1 immediately generalizes to the multiclass setting and the proof requires no modification, i.e,

$$(\widehat{A}^{\lambda}_{1}, \ldots, \widehat{A}^{\lambda}_{M}) \Rightarrow (\sigma_1 B_1, \ldots, \sigma_M B_M) \quad \text{in } \mathcal{D}^M \text{ as } \lambda \to \infty, \tag{EC.57}$$

where $B_1, \ldots, B_M$ are M independent standard Brownian motions, and $\sigma^2_m = 1 + \nu_m$, $m \in \mathcal{M}$.

The asymptotic independence in Theorem 4.3 continues to hold in the multi-class setting for each pair of in-house call center and outsourcer. The proof of that theorem reveals that the only requirement is that the processes in station O are $\mathcal{C}$-tight. (See also the intuition in Remark 4.3.) In the multi-class settings, such tightness holds, in particular, if the sequence $\{(\widehat{X}^{\lambda}_{1,O}, \ldots \widehat{X}^{\lambda}_{M,O})\}$ converges to a continuous limit. This is satisfied, for example, by the QIR family of controls studied in Gurvich and Whitt [2009]. The same is true for other policies that try to maintain certain proportions between the queues; see Atar [2005] and Atar et al. [2004]. However, not all routing rules satisfy this property. For example, the bang-bang type rule in Harrison and Zeevi [2004] does not produce continuous heavy-traffic limits.

If the $\mathcal{C}$-tightness holds, then the corresponding result is a direct extension of Theorem 4.3 – requiring

again no modification of the proof, as the analysis can be applied to each $m \in \mathcal{M}$ separately. The corresponding asymptotic independence statement is then

$$\mathbb{P}\left\{\widehat{X}^{\lambda}_{m,O}(t) > q, D^{\lambda}_{m,I}(t) = d\right\} = \mathbb{P}\left\{\widehat{X}^{\lambda}_{m,O}(t) > q\right\}\mathbb{P}\left\{D^{\lambda}_{m,I}(t) = d\right\} + o(1),\ t > 0, \quad q \in \mathbb{R}, d \in \mathbb{Z}_+,$$

where $D^{\lambda}_{m,I} := N^{\lambda}_m + K^{\lambda}_m - X^{\lambda}_{m,I}$. Corollaries 4.4 and 4.5 also extend to the multi-class case without any changes to their proofs.

EC.2.1. Optimization in the Multi-class Setting

As we pointed out in the introduction, one may wish to study the value of real-time information about the state of the in-house call centers for optimization problems of the outside provider. We will now show that such information carries at most marginal benefit. For ease of presentation we focus on a relatively simple setting, but the results hold for more general models. Specifically, we focus on the case in which the outside provider has a single pool of servers serving all customer classes (overflow streams). A two-class example is depicted in Figure 1(b). The queueing network in station O is then the so-called V model; see Gurvich et al. [2008] and Atar et al. [2004].

In serving multiple input streams in one facility, the outside provider's problem is now a staffing-and-routing optimization problem. Specifically, the outside provider's problem (5) may become

$$\begin{aligned} &\min\ C_s(N_O) + \sum_{m\in\mathcal{M}} \mathbb{E}\left[\int_0^{\infty} e^{-\gamma s} C_m(\widehat{Q}^{\pi,\lambda}_{m,O}(s))ds\right] \\ &\text{s.t.}\quad N_O \in \mathbb{Z}_+, \pi \in \Pi^{\lambda}, \end{aligned} \tag{EC.58}$$

where Π^{λ} is the family of admissible routing rules, and the superscript π makes explicit the dependency of the queue length process on the control being employed. Normally, as in (5), the outside provider would have a constraint on the waiting time rather than holding costs, but one may think about that problem as a dualization of such constraints. In fact, it suffices to fix N_O and consider the control problem

$$\min \sum_{m\in\mathcal{M}} \mathbb{E}\left[\int_0^{\infty} e^{-\gamma s} C_m(\widehat{Q}^{\pi,\lambda}_{m,O}(s))ds\right]$$

$$\text{s.t.} \quad \pi \in \Pi^{\lambda}. \tag{EC.59}$$

Assuming that (EC.58) has an optimal staffing-and-routing solution $(N_O^{*,\lambda}, \pi^{*,\lambda})$ and that, for each N_O, (EC.59) has an optimal control $\pi^{*,\lambda}(N_O)$, one can solve (EC.58) by using $\pi^{*,\lambda}(N_O)$ and optimizing only over N_O.

We further assume that the functions $C_m(\cdot)$ are twice continuously differentiable with strictly positive second derivatives, and that these functions have at most polynomial growth, i.e, that there exist constant M_m and l_m such that $C_m(x) \leq M_m(1+|x|^{l_m})$. These assumptions will allow us to interpret directly some results in Atar et al. [2004].

To show that the real-time state information does not carry significant value we first formalize the notion of information with respect to the optimization problem (EC.59). To that end, we define $S^{\lambda}_{m,O}(t)$ to be the number of class-m customers that departed from station O by time t after completing service, and $L^{\lambda}_{m,O}(t)$ be the number of class-m customers that abandoned station O by time t. We assume that, regardless of whether information is shared or not, the outside provider knows the state of his local queues, as reflected in the processes $X^{\lambda}_{m,O}$ and $Q^{\lambda}_{m,O}$, as well as the local history up to the time of the decision. Namely, when making a decision at time t, the outside provider knows the evolution of the processes $(X^{\lambda}_{m,O}, Q^{\lambda}_{m,O})$, $m \in \mathcal{M}$ up to time t, as well as that of the processes $A^{\lambda}_{m,O}$, $S^{\lambda}_{m,O}$ and $L^{\lambda}_{m,O}$.

The information-sharing and no-information-sharing settings differ with regards to whether or not the outside provider has access to the real-time state information of each of the in-house queues as captured by the process $(X^{\lambda}_{1,I}, \ldots, X^{\lambda}_{M,I})$. Without information sharing the outside provider has to make its prioritization decisions based only on his local information. Define the filtrations

$$\mathcal{F}^{\lambda}_t = \sigma\Big\{A^{\lambda}_{m,O}(s), S^{\lambda}_{m,O}(s), L^{\lambda}_{m,O}(s), X^{\lambda}_{m,O}(s), Q^{\lambda}_{m,O}(s), Z^{\lambda}_{m,O}(s), X^{\lambda}_{m,I}(s); m \in \mathcal{M}, s \leq t\Big\}.$$

and

$$\check{\mathcal{F}}^{\lambda}_t = \sigma\Big\{A^{n}_{m,O}(s), S^{\lambda}_{m,O}(s), L^{\lambda}_{m,O}(s), X^{\lambda}_{m,O}(s), Q^{\lambda}_{m,O}(s), Z^{\lambda}_{m,O}(s); m \in \mathcal{M}, s \leq t\Big\}.$$

Note that $\mathcal{F}^{\lambda}_t$ contains also the information about $X^{\lambda}_{m,I}$ which is not contained in $\check{\mathcal{F}}^{\lambda}_t$. We then let Π^{λ} be the family of policies that are "adapted" to $(\mathcal{F}^{\lambda}_t)_{t\geq 0}$, and $\check{\Pi}^{\lambda}$ be the subset of Π^{λ} that is adapted to $(\check{\mathcal{F}}^{\lambda})_{t\geq 0}$.

We refer the reader to Atar et al. [2004] for a more formal discussion. The control problem for the outside provider can then be defined by

$$\begin{aligned} \min \ & \sum_{m\in\mathcal{M}} \mathbb{E}\left[\int_0^\infty e^{-\gamma s} C_m(\widehat{Q}^{\pi,\lambda}_{m,O}(s))ds\right] \\ \text{s.t.} \ & \check{\pi}\in\check{\Pi}^\lambda, \end{aligned} \tag{EC.60}$$

and this should be contrasted with (EC.59), where the larger family of policies Π^λ is admissible. Let $C^\lambda(\pi,x)$ be the cost under policy π when the initial state is x. By state we mean the state throughout the system (in-house stations and outside provider). Let $V^\lambda(x)=\inf_{\pi\in\Pi^\lambda} C^\lambda(\pi,x)$ and $\check{V}^\lambda(x)=\inf_{\check{\pi}\in\check{\Pi}^\lambda} C^\lambda(\check{\pi},x)$ be the optimal costs to (EC.59) and (EC.60), respectively.

One expects the optimal value of (EC.60) to be strictly greater than that of (EC.59), i.e, that $V^\lambda(x)<\check{V}^\lambda(x)$. The main result of this section shows that the reverse inequality also holds asymptotically. Namely, that in a context with resource pooling, the additional information about the in-house call centers does not carry significant benefits.

THEOREM EC.2.1. Assume that (EC.55) holds. Let $x^\lambda=(\widehat{Q}^\lambda_{m,O}(0),\widehat{X}^\lambda_{m,O}(0),\widehat{X}^\lambda_{m,I}(0):m\in\mathcal{M})$ and suppose that the sequence $\{x^\lambda\}$ satisfies (EC.56) where the limit $x=\lim_{\lambda\to\infty}x^\lambda$ is nonrandom. Then,

$$\check{V}^\lambda(x^\lambda)\le V^\lambda(x^\lambda)+o(1).$$

Our approximation of the overflow process in (EC.57) is key in establishing the above result. In fact, given that approximation, the proof of Theorem EC.2.1 (see below) is a direct corollary of the analysis and the results in Atar et al. [2004]. Theorem EC.2.1 adds to our discussion of outsourcing-scheme comparisons in §3.1, where we argued how our asymptotic-independence results allow to simplify the outsourcer problem to one that does not require knowledge of joint distributions. Here we show that even real-time information carries, at most, negligible benefit for the outside provider.

Outline of the proof: We provide only an outline of the argument, assuming familiarity of the reader with the terminology and notation in Atar et al. [2004]. First note that, due to the convergence of the overflow processes (which would be the input processes in the model of Atar et al. [2004]), one can follow the steps in

§2.5 there to "guess" the Brownian control problem. This Brownian control problem should be interpreted in our setting as focusing on the V model in station O. Hence, it uses only local information about station O as we do in defining the family $\check{\Pi}^{\lambda}$ of policies. Due to our restrictions on the cost functions $C_m(\cdot),\ m \in \mathcal{M}$, Assumptions 1-3 of Atar et al. [2004] hold. In turn, the requirements of Theorem 1 there are satisfied, and the Brownian control problem has an optimal Markov control policy. In fact, due to our restrictions on the cost functions, it also follows from Proposition 3 there that the Markov control function h is locally Hölder continuous (see the statement of Theorem 2 there). Atar et al. [2004] then uses this Markov control in proposing a sequence of controls for the original sequence of queueing system (see §2.6 there).

By Theorem 2 in Atar et al. [2004] it follows that the policies constructed there are asymptotically optimal for (EC.60) that we defined. It remains to argue that these are actually asymptotically optimal for (EC.59). In our setting this follows from our convergence results for the overflow process (see Theorem 4.1) and repeating precisely the steps in the proof of asymptotic optimality in §4 of Atar et al. [2004]. In following that proof, one should note that, given the convergence of the overflow processes to a drifted Brownian motion that is independent of the in-house processes, the arguments remain unchanged when the larger information set is being employed. ■

e-companion references

Asmussen, S. 2003. *Applied Probability and Queues*. 2nd ed. Springer, New York.

Atar, R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Annals of Applied Probability* **15**(4) 2606–2650.

Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multi-class queue with many exponential servers: asymptotic optimality in heavy traffic. *Annals of Applied Probability* **14**(3) 1084–1134.

Billingsley, P. 1968. *Convergence of Probability Measures*. J. Wiley & Sons, New York.

Glynn, P.W., W. Whitt. 1993. Limit theorems for cumulative processes. *Stochastic Processes and their Applications* **47** 299–314.

Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service-level differentiation in call centers with fully flexible servers. *Management Science* **54**(2) 279–294.

Gurvich, I., W. Whitt. 2009. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research* **34**(2) 363–396.

Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. *Operations Research* **58**(2) 316–328.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime. *Operations Research* **52**(2) 243–257.

Iglehart, D.L., W. Whitt. 1971. The equivalence of functional central limit theorems for counting processes and associated partial sums. *The Annals of Mathematical Statistics* **42**(4) 1372–1378.

Jakubowski, A. 1996. Convergence in various topologies for stochastic integrals driven by semimartingales. *The Annals of Probability* **24**(4) 2141–2153.

Karatzas, I., S. Shreve. 1991. *Brownian Motion and Stochastic Ccalculus*. 2nd ed. Springer-Verlag, New York.

Meyn, S.P. 2008. *Control techniques for complex networks*. Cambridge Univ Pr.

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

Perry, O., W. Whitt. 2010a. A fluid limit for an overloaded x model via an averaging principle. Working paper, Columbia University, New York, NY.

Perry, O., W. Whitt. 2010b. Gaussian approximations for an overloaded x model via and averaging principle. Working paper, Columbia University, New York, NY.

Ross, S.M. 1996. *Stochastic processes*. Wiley New York.

Talreja, R., W. Whitt. 2009. Heavy-traffic limits for waiting times in many-server queues with abandonment. *Annals of Applied Probability* **19**(6) 2137–2175.

Whitt, W. 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* **37**(3) 307–314.

Whitt, W. 2002. *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer Series in Operations Research, New York.

Whitt, W. 2005. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research* **30**(1) 1–27.