# Electronic Companion

## EC.1. The $G/M/\infty$ queue length — a boundary-free example

The $G/M/\infty$ system provides an instructive example that features both general and exponential clocks, and in contrast to the $G/G/1$ setting, the diffusion error for the infinite-server queue is composed solely of interior terms and no boundary terms. As a result, obtaining a three-moment error bound is considerably simpler.

The $G/M/\infty$ system is an infinite-server queue where arrivals form a renewal process with generic interarrival time distribution $U$ and exponentially distributed service times with rate $\mu$. Let $R=\lambda/\mu$ be the offered load. We assume that $\mathbb{E}U^3<\infty$ and set

$$\lambda = 1/\mathbb{E}U, \quad c_U^2 = \lambda^2 \mathrm{Var}(U).$$

Let $Q(t)$ be the customer count, let $R_a(t)$ be the residual interarrival time at $t\geq 0$, and assume that

$$\{Z(t) = (Q(t), R_a(t))\}$$

is positive Harris recurrent. Let $Z=(Q,R_a)$ denote the vector having the stationary distribution. Let $\delta = 1/\sqrt{R}$ and define the normalized customer count

$$X(t) = \delta(Q(t)-R) \quad \text{and} \quad X = \delta(Q-R).$$

The following is the main result of this section.

Theorem EC.1. *Let $Y\sim N(0,\sigma^2)$ with $\sigma^2=(1+c_U^2)$. Then*

$$\begin{aligned}
&d_W(X,Y)\leq \delta(K_A+K_D),\\
&K_A = \frac{1}{3(1+c_U^2)}\mathbb{E}\left|1-\lambda U\right|^3 + \frac{c_U^2}{1+c_U^2}\Big(1+\frac{1}{2}\lambda^2\mathbb{E}U^2(\delta^4\lambda^3\mathbb{E}U^3/3+\delta^2\lambda^2\mathbb{E}U^2+1)\Big),\\
&K_D = \sqrt{\frac{2/\pi}{(1+c_U^2)}}\sqrt{\mathbb{E}X^2}\sqrt{\lambda^3\mathbb{E}U^3/3} + \frac{1}{(1+c_U^2)}\Big(\delta^2\lambda^3\mathbb{E}U^3/3+\lambda^2\mathbb{E}U^2+\frac{1}{3}\Big).
\end{aligned}$$

The proof of Theorem EC.1 follows the same structure as the proof of Theorem 1. Namely, we use the BAR for the compensated customer count in the infinite-server queue to extract a diffusion generator and corresponding error terms in Proposition EC.1. Bounding the error terms requires both Stein factor bounds and a bound on the second moment of the centered queue length (unlike our $G/G/1$ example, which did not require moment bounds). The former are well known and the latter is obtained from the BAR by using a quadratic test function. After presenting said Stein factor and moment bounds, we combine everything to prove Theorem EC.1 at the end of the section.

For the BAR, let $A(t)$ and $U(t)$ be defined as in Section 2.1, and let $D(t)$ denote the number of departures on $[0,t]$. Given a function $f(Z(t))$, partition $[0,t]$ according to event times and consider the FTC conditions analogous to those in Section 2.2, as well as the integrability condition

$$\mathbb{E}\left|f(Z)\right|,\ \mathbb{E}\left|\partial_{r_a} f(Z)\right|,\ \mathbb{E}\int_0^t \left|\Delta f(Z(s-))\right| dD(s),\ \mathbb{E}\int_0^t \left|\Delta f(Z(s-))\right| dA(s) < \infty.$$

Arguing as in Section 2.2 yields the BAR.

LEMMA EC.1. *Initialize $Z(0)\sim Z$. If $f(Z(s))$ satisfies the FTC conditions with probability one under $Z(0)\sim Z$ and if the integrability condition holds, then for all $t\geq 0$,*

$$\begin{aligned} 0 =&\ -t\mathbb{E}(\partial_{r_a} f(Z)) + \mathbb{E}\int_0^t \Delta f(Z(s-))dD(s) + \mathbb{E}\int_0^t \Delta f(Z(s-))dA(s) \\ =&\ -t\mathbb{E}(\partial_{r_a} f(Z)) + t\mu\mathbb{E}Q(f(Q-1,R_a) - f(Q,R_a)) + \mathbb{E}\int_0^t \Delta f(Z(s-))dA(s), \end{aligned} \tag{EC.1}$$

*provided that all expectations on the right-hand side are well defined (in a manner similar to (3)).*

*Proof of Lemma EC.1* The first equality follows as in Lemma 1 and the second is due to Theorem 3.3.1 of Baccelli and Brémaud (2003). □

REMARK EC.1. For convenience, we will work with the right-hand side of (EC.1) with $t=1$ there. For an approach that avoids using the machinery of Baccelli and Brémaud

(2003) to establish the second equality in (EC.1), consider the first equality in (EC.1), divide both sides by $t$, and let $t \to 0$ to get

$$\begin{aligned}0 = & -\mathbb{E}(\partial_{r_a} f(Z)) + \lim_{t\to 0}\frac{1}{t}\mathbb{E}\int_0^t \Delta f(Z(s-))dD(s) + \lim_{t\to 0}\frac{1}{t}\mathbb{E}\int_0^t \Delta f(Z(s-))dA(s)\\ = & -\mathbb{E}(\partial_{r_a} f(Z)) + \mu\mathbb{E}Q(f(Q-1,R_a) - f(Q,R_a)) + \lim_{t\to 0}\frac{1}{t}\mathbb{E}\int_0^t \Delta f(Z(s-))dA(s).\end{aligned}$$

The second equality is argued using the fact that $\mathbb{P}(D(t)=1) = t\mu Q(0) + o(t)$ and $\mathbb{P}(D(t) > 1) = o(t)$ for $t$. Since $\mathbb{E}A(t) = \lambda t$, the presence of the limit as $t \to 0$ in the third term does not add further complexity (compared to the $t = 1$ case) to our diffusion generator extraction procedure (Proposition EC.1).

Defining the compensated customer count

$$\widetilde{X}(t) = X(t) - \delta\lambda R_a(t) \quad \text{and} \quad \widetilde{X} = X - \delta\lambda R_a, \tag{EC.2}$$

and specializing the BAR (EC.1) to $\widetilde{X}$ yields

$$\delta\lambda\mathbb{E}f'(\widetilde{X}) + \mu\mathbb{E}\big(Q\big(f(\widetilde{X}-\delta) - f(\widetilde{X})\big)\big) + \mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dA(t) = 0. \tag{EC.3}$$

The following expansion of (EC.3) is analogous to Proposition 1 and is proved in Appendix EC.6.

Proposition EC.1. *If $f \in C^2(\mathbb{R})$ with $f''(x)$ absolutely continuous and $\|f''\|, \|f'''\| < \infty$, then, provided that all expectations are well defined,*

$$\mu\mathbb{E}\big(Q\big(f(\widetilde{X}-\delta) - f(\widetilde{X})\big)\big) = -\delta R f'(\widetilde{X}) - \mu\mathbb{E}Xf'(X) + \frac{1}{2}\mu\mathbb{E}f''(X) + \epsilon_D(f), \tag{EC.4}$$

$$\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dA(t) = \frac{1}{2}\mu c_U^2\mathbb{E}f''(X) + \epsilon_A(f) \tag{EC.5}$$

*where*

$$\begin{aligned}|\epsilon_A(f)| &\le \frac{1}{6}\delta\mu\|f'''\|\mathbb{E}\,|1-\lambda U|^3 + \frac{1}{2}\delta\mu c_U^2\|f'''\|\Big(1 + \frac{1}{2}\lambda^2\mathbb{E}U^2(\mathbb{E}(Q-R)^2/R^2 + 1)\Big),\\ |\epsilon_D(f)| &\le \mu\delta\|f''\|\sqrt{\mathbb{E}X^2}\sqrt{\lambda^3\mathbb{E}U^3/3} + \frac{1}{2}\mu\delta\|f'''\|\delta^2\lambda\mathbb{E}(QR_a) + \frac{1}{6}\mu\delta\|f'''\|.\end{aligned}$$

The upper bounds on $|\epsilon_A(f_h)|$ and $|\epsilon_D(f_h)|$ contain terms involving $\mathbb{E}(Q-R)^2$ and $\mathbb{E}(QR_a)$. The following lemma presents some useful identities, including a bound on these quantities. It is proved in Appendix EC.6.1.

Lemma EC.2. *Recall that $R=\lambda/\mu$. For any $m>1$,*

$$\mathbb{E}A(1)=\lambda, \quad \mathbb{E}R_a^{m-1}=\lambda\mathbb{E}U^m/m, \quad \mathbb{E}Q=R, \tag{EC.6}$$

$$\mathbb{E}(Q-R)^2=\lambda\mathbb{E}QR_a\le\lambda^3\mathbb{E}U^3/3+R\lambda^2\mathbb{E}U^2, \tag{EC.7}$$

$$\mathbb{E}\int_0^1 Q(t-)dA(t)=\mu\mathbb{E}(Q-R)^2+\lambda R-\lambda. \tag{EC.8}$$

Applying the expansions in (EC.4)–(EC.5) to the BAR (EC.3) suggests the diffusion approximation with generator

$$G_Y f(x)=-\mu x f'(x)+\frac{1}{2}\mu(1+c_U^2)f''(x), \quad x\in\mathbb{R},$$

which corresponds to the $N(0,\sigma^2)$ distribution with $\sigma^2=(1+c_U^2)$. The following lemma is a rescaled version of (Chen et al. 2011, Lemma 2.4), and is proved in Appendix EC.6.2.

Lemma EC.3. *Let $Y\sim N(0,\sigma^2)$ and let $f_{h,\sigma}(x)$ be the solution to*

$$-xf'_{h,\sigma}(x)+\frac{1}{2}\sigma^2 f''_{h,\sigma}(x)=\frac{1}{\mu}(\mathbb{E}h(Y)-h(x)), \quad x\in\mathbb{R}, \tag{EC.9}$$

*with $f_{h,\sigma}(0)=0$. Then*

$$\|f'_{h,\sigma}\|\le\frac{2}{\mu}, \quad \|f''_{h,\sigma}\|\le\frac{\sqrt{2/\pi}}{\mu\sigma}, \quad \text{and} \quad \|f'''_{h,\sigma}\|\le\frac{2}{\mu\sigma^2}.$$

*As a consequence, $|f_{h,\sigma}(x)|\le 2|x|/\mu$ for all $x\in\mathbb{R}$.*

We are ready to prove Theorem EC.1.

*Proof of Theorem EC.1* Since $\|f_h'\|<\infty$ and $|f_h(x)|\le 2|x|/\mu$ for any $h\in\mathrm{Lip}(1)$, one readily verifies that the BAR (EC.3) holds with $f_h(X-\delta\lambda R_a)$. Applying the Stein factor bounds of Lemma EC.3 and the moment bound in (EC.7) of Lemma EC.2 to the upper bounds on $|\epsilon_A(f_h)|$ and $|\epsilon_D(f_h)|$ in Proposition EC.1 yields $|\epsilon_A(f_h)|\le K_A$ and $|\epsilon_D(f_h)|\le K_D$. □

The following appendix treats the two-station tandem queue, the only example whose diffusion approximation is multidimensional.

## EC.2. The tandem queue — a multidimensional example

In higher dimensions, the mechanics are unchanged: we derive the diffusion generator from the BAR and decompose the approximation error into interior and boundary contributions. Our approach also extends naturally to generalized Jackson networks. The added difficulty lies in establishing Stein factor bounds for the multidimensional RBMs; bounds are not known even for the two-dimensional RBM approximating the tandem queue.

Consider two first-come-first-served single-server stations in tandem. Customers arrive to station one according to a renewal process with generic interarrival time $U$; after service at station one they move to station two and depart after completing service at station two. Service times at station $i$ are i.i.d. $S_i$, independent of the arrival process and of service times at the other station. Assume $\mathbb{E}U^3 < \infty$, $\mathbb{E}S_1^3, \mathbb{E}S_2^3 < \infty$, and no simultaneous events almost surely. Set

$$\lambda = 1/\mathbb{E}U, \quad \mu_i = 1/\mathbb{E}S_i, \quad \rho_i = \lambda/\mu_i,$$
$$c_U^2 = \lambda^2 \mathrm{Var}(U), \quad c_{S,i}^2 = \mu_i^2 \mathrm{Var}(S_i).$$

For $t \geq 0$, let $Q_i(t)$ be the queue length at station $i$, $R_a(t)$ the residual interarrival time, and $R_{s,i}(t)$ the remaining service at station $i$ (if $Q_i(t) = 0$, the service time of the next customer). Define $X(t) = (\delta_1 Q_1(t), \delta_2 Q_2(t))$ with $\delta_i = 1 - \rho_i$, $R_s(t) = (R_{s,1}(t), R_{s,2}(t))$, and $Z(t) = (X(t), R_a(t), R_s(t))$. Assume $\{Z(t)\}$ is positive Harris recurrent (cf. Dai and Meyn (1995)) and let $Z = (X, R_a, R_s)$ have the stationary distribution. Let $A(t)$ and $D_i(t)$ denote arrivals and station-$i$ departures on $[0, t]$.

The procedure of applying the generator approach to the tandem queue is similar to the three examples already considered. We therefore highlight only the key steps, leaving the formal derivations to the interested reader. For sufficiently regular functions, the BAR for $Z$ is

$$\begin{aligned} &- \mathbb{E}(\partial_{r_a} f(Z)) - \sum_{i=1}^{2} \mathbb{E}(1(X_i > 0)\partial_{r_{s,i}} f(Z)) \\ &+ \mathbb{E}\int_0^1 \Delta f(Z(t-))dA(t) + \sum_{i=1}^{2} \mathbb{E}\int_0^1 \Delta f(Z(t-))dD_i(t) = 0. \end{aligned} \tag{EC.10}$$

Let $\widetilde{X}$ have the stationary distribution of the compensated queue-length vector $\widetilde{X}(t) = (\widetilde{X}_1(t), \widetilde{X}_2(t))$, where

$$\widetilde{X}_1(t) = X_1(t) - \delta_1 \lambda R_a(t) + \delta_1 \mu_1 R_{s,1}(t), \quad \widetilde{X}_2(t) = X_2(t) - \delta_2 \mu_1 R_{s,1}(t) + \delta_2 \mu_2 R_{s,2}(t).$$

The BAR for $\widetilde{X}$ is

$$\begin{aligned}
&\delta_1 \mathbb{E}\big((\lambda - \mu_1 1(Q_1 > 0))\partial_{x_1} f(\widetilde{X})\big) + \delta_2 \mathbb{E}\big((\mu_1 1(Q_1 > 0) - \mu_2 1(Q_2 > 0))\partial_{x_2} f(\widetilde{X})\big) \\
&+ \mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dA(t) + \sum_{i=1}^{2} \mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dD_i(t) = 0,
\end{aligned}$$

Omitting the higher-order error terms, it follows that

$$\begin{aligned}
&\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dA(t) \approx \frac{1}{2}\delta_1^2 \lambda c_U^2 \mathbb{E}\partial_{x_1}^2 f(X), \\
&\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dD_1(t) \approx \frac{1}{2}\mu_1 c_{S,1}^2 \mathbb{E}\big(\delta_1^2 \partial_{x_1}^2 f(X) - 2\delta_1\delta_2 \partial_{x_1}\partial_{x_2} f(X) + \delta_2^2 \partial_{x_2}^2 f(X)\big), \\
&\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dD_2(t) \approx \frac{1}{2}\delta_2^2 \mu_2 c_{S,2}^2 \mathbb{E}\partial_{x_2}^2 f(X).
\end{aligned}$$

As in our prior examples, the omitted error terms can be bounded using the first three moments of the primitives as well as the second- and third-order derivatives of $f(x)$. Furthermore, omitting the boundary error terms, we have

$$\begin{aligned}
&\delta_1 \mathbb{E}\big((\lambda - \mu_1 1(Q_1 > 0))\partial_{x_1} f(\widetilde{X})\big) + \delta_2 \mathbb{E}\big((\mu_1 1(Q_1 > 0) - \mu_2 1(Q_2 > 0))\partial_{x_2} f(\widetilde{X})\big) \\
\approx\ & -\mu_1 \delta_1^2 \mathbb{E}\partial_{x_1} f(X) + \delta_2(\mu_1\delta_1 - \mu_2\delta_2)\mathbb{E}\partial_{x_2} f(X) \\
&+ \mu_1\big(\delta_1 1(Q_1 = 0)\partial_{x_1} f(X) - \delta_2 1(Q_1 = 0)\partial_{x_2} f(X)\big) + \mu_2 \delta_2 1(Q_2 = 0)\partial_{x_2} f(X).
\end{aligned}$$

The second line on the right-hand side is tied to the reflection structure of the approximating RBM, which we introduce shortly. It is similar to the $f'(0)$ found in (8) of Proposition 1 and is not simply an error term.

Our expansions suggest the diffusion generator

$$\begin{aligned}
G_Y f(x) =\ & -\mu_1 \delta_1^2 \partial_{x_1} f(x) + \delta_2(\mu_1\delta_1 - \mu_2\delta_2)\partial_{x_2} f(x) + \frac{1}{2}\delta_1^2\big(\lambda c_U^2 + \mu_1 c_{S,1}^2\big)\partial_{x_1}^2 f(x) \\
& - \delta_1\delta_2 \mu_1 c_{S,1}^2 \partial_{x_1}\partial_{x_2} f(x) + \frac{1}{2}\delta_2^2\big(\mu_1 c_{S,1}^2 + \mu_2 c_{S,2}^2\big)\partial_{x_2}^2 f(x) \\
& + \mu_1(1(x_1 = 0)(\delta_1 \partial_{x_1} f(x) - \delta_2 \partial_{x_2} f(x))) + \mu_2 \delta_2 1(x_2 = 0)\partial_{x_2} f(x), \quad x \in \mathbb{R}_+^2,
\end{aligned}$$

which corresponds to a two-dimensional RBM $\{Y(t) \in \mathbb{R}^2_+ : t \geq 0\}$ on the nonnegative orthant defined as follows. Set

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \lambda c_U^2 + \mu_1 c_{S,1}^2 & -\mu_1 c_{S,1}^2 \\ -\mu_1 c_{S,1}^2 & \mu_1 c_{S,1}^2 + \mu_2 c_{S,2}^2 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix},$$

let $\{\xi(t) : t \geq 0\}$ be a two-dimensional Brownian motion with drift $b = -R\mu$ and covariance matrix $\Sigma$, and let

$$Y(t) = \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix} \widetilde{Y}(t),$$

where

$$\widetilde{Y}(t) = \xi(t) + RI(t), \quad t \geq 0,$$

and $I : R \to \mathbb{R}^2$ is the unique nondecreasing process with $I(0) = 0$ and

$$\int_0^\infty \widetilde{Y}_i(t) dI_i(t) = 0, \quad i = 1, 2.$$

For the existence and uniqueness of $\{\widetilde{Y}(t) : t \geq 0\}$ see Harrison and Reiman (1981).

The final ingredient is the Poisson equation. Fix $h : \mathbb{R}^2 \to \mathbb{R}$ with $\mathbb{E}|h(Y)| < \infty$, where $Y$ has the stationary distribution of $\{Y(t) : t \geq 0\}$, and consider

$$G_Y f_h(x) = \mathbb{E}h(Y) - h(x), \qquad x \in \mathbb{R}^2_+,$$

$$\delta_1 \partial_{x_1} f(x) - \delta_2 \partial_{x_2} f(x) = 0, \qquad x = (0, x_2) \in \mathbb{R}^2_+, \tag{EC.11}$$

$$\partial_{x_2} f(x) = 0, \qquad x = (x_1, 0) \in \mathbb{R}^2_+. \tag{EC.12}$$

This is known as an oblique derivative problem (Lieberman 2013) with (EC.11) and (EC.12) arising from the reflection structure of the RBM, which is driven by the matrix $R$. To bound the diffusion approximation error for the tandem queue, we require bounds on the partial derivatives, up to the third order, of $f_h(x)$. These Stein factor bounds remain an open problem.

## EC.3. The $G/G/1$ workload: supporting proofs

### EC.3.1. Proof of the inversion formula (Lemma 3)

Let $\tau_m$ be the time of the $m$th arrival and observe that

$$\int_0^1 f(X(t))dt = \int_0^{\tau_1} f(X(t))dt + \sum_{m=1}^{A(1)} \int_{\tau_m}^{\tau_{m+1}} f(X(t))dt - \int_1^{\tau_{A(1)+1}} f(X(t))dt.$$

Since $\tau_{m+1} - \tau_m = U(\tau_m)$, it follows that

$$\begin{aligned}\sum_{m=1}^{A(1)} \int_{\tau_m}^{\tau_{m+1}} f(X(t))dt &= \sum_{m=1}^{A(1)} \int_0^{U(\tau_m)} f\big(X(\tau_m + u)\big)du \\ &= \int_0^1 \int_0^{U(t)} f\big(X(t+u)\big)dudA(t).\end{aligned}$$

Furthermore, since $\tau_1 = R_a(0)$ and $\tau_{A(1)+1} = 1 + R_a(1)$,

$$\int_0^{\tau_1} f(X(t))dt - \int_1^{\tau_{A(1)+1}} f(X(t))dt = \int_0^{R_a(0)} f(X(t))dt - \int_0^{R_a(1)} f(X(1+t))dt.$$

Therefore,

$$\begin{aligned}\mathbb{E}f(X) &= \mathbb{E}\int_0^1 f(X(t))dt \\ &= \mathbb{E}\int_0^1 \int_0^{U(t)} f\big(X(t+u)\big)dudA(t) + \mathbb{E}\bigg(\int_0^{R_a(0)} f(X(t))dt - \int_0^{R_a(1)} f(X(1+t))dt\bigg)\end{aligned}$$

and the result follows by noting that since $(X, R_a)$ is stationary, the joint distribution of $(X(t), R_a(t))$ is the same as $(X(1+t), R_a(1+t))$, so the two integrals cancel in expectation. □

### EC.3.2. Proof of Theorem 1

Given $h \in \text{Lip}(1)$, consider the Poisson equation (12) with $\theta = \delta^2$ and $\sigma^2 = \delta^2 \rho \mathbb{E}S(c_U^2 + c_S^2)$. Assuming that the BAR (7) holds for $f_h(\widetilde{X})$, we set $x = X$ and take expected values to arrive at

$$\begin{aligned}\mathbb{E}h(Y) - \mathbb{E}h(X) &= \mathbb{E}G_Y f_h(X) \\ &= \mathbb{E}G_Y f_h(X) - \delta\mathbb{E}\big((\rho - 1(X>0))f_h'(\widetilde{X})\big) - \mathbb{E}\int_0^1 \Delta f_h(\widetilde{X}(t-))dA(t).\end{aligned}$$

Taking the supremum over all $h \in \mathrm{Lip}(1)$ and using Proposition 1 yields

$$d_W(X,Y) \le |\epsilon_0(f_h)| + |\epsilon_A(f_h)|.$$

We now bound both $|\epsilon_0(f_h)|$ and $|\epsilon_A(f_h)|$, and then show that the BAR (7) holds for $f_h(\widetilde{X})$. Recalling that

$$\epsilon_0(f_h) = \delta^3 \rho \mathbb{E}(R_a f_h''(\xi)) - \delta^2 \rho \mathbb{E}(1(X=0) R_a f_h''(\xi)),$$

the bound on $|\epsilon_0(f_h)|$ follows from $\|f_h''\| \le \delta^{-2}$ (Lemma 4) and $\mathbb{E}R_a = \lambda \mathbb{E}U^2/2$ (Lemma 2). Similarly, recall that

$$\begin{aligned}\epsilon_A(f_h) = &\frac{1}{6}\delta^3 \mathbb{E}\int_0^1 (S(t) - \rho U(t))^3 f_h'''(\xi(t)) dA(t) \\ &- \frac{1}{2}\delta^2 \lambda \mathbb{E}(S-\rho U)^2 \mathbb{E}\int_0^1 \int_0^{U(t)} \big(X(t+u) - X(t-)\big) f_h'''(\xi(t+u)) du dA(t).\end{aligned}$$

The first term is handled using the Stein factor bound $\|f_h'''\| \le 4/\sigma^2$, while the second term requires controlling increments of $X$ over an interarrival time. Namely,

$$\frac{1}{6}\delta^3 \mathbb{E}\left|\int_0^1 (S(t) - \rho U(t))^3 f_h'''(\xi(t)) dA(t)\right| \le \frac{4\delta^3 \mathbb{E}|S-\rho U|^3}{6\delta^2 \rho \mathbb{E}S(c_U^2 + c_S^2)} \mathbb{E}A(1) = \frac{2\delta \mathbb{E}|S-\rho U|^3}{3(\mathbb{E}S)^2(c_U^2+c_S^2)},$$

where the equality is true because $\mathbb{E}A(1) = \lambda$, and

$$\begin{aligned}&\frac{1}{2}\delta^2 \lambda \mathbb{E}(S-\rho U)^2 \mathbb{E}\left|\int_0^1 \int_0^{U(t)} \big(X(t+u) - X(t-)\big) f_h'''(\xi(t+u)) du dA(t)\right| \\ &\le 2\mathbb{E}\int_0^1 \int_0^{U(t)} |X(t+u) - X(t-)| \, du dA(t) \\ &= 2\mathbb{E}\int_0^1 \int_0^{U(t)} \big|(X(t-) + \delta S(t) - \delta u)^+ - X(t-)\big| \, du dA(t) \\ &\le 2\mathbb{E}\int_0^1 \int_0^{U(t)} \delta(S(t) + U(t)) du dA(t) \\ &= 2\delta \mathbb{E}(US + U^2)\mathbb{E}A(1) = 2\delta(\mathbb{E}S + \lambda \mathbb{E}U^2),\end{aligned}$$

where the last inequality follows from $|(x+y)^+ - x^+| \le y^+$.

It remains to show that (7) holds with $f(\widetilde{X}) = f_h(\widetilde{X})$. By assumption (1), $U$ and $S$ have finite third moments. Since $\mathbb{E}R_a = \lambda \mathbb{E}U^2/2 < \infty$ (Lemma 2) and $\mathbb{E}X^2 < \infty$ (Asmussen 2003, Theorems X.3.4 and X.2.1), the fact that $|f_h(x)| \le x^2/(2\theta)$ (Lemma 4) yields $\mathbb{E}|f_h(\widetilde{X})| < \infty$; the fact that $\mathbb{E}|f_h'(\widetilde{X})| < \infty$ is argued similarly. To show that the jump term in (7) is well defined, note that

$$\begin{aligned}
\mathbb{E}\left|\int_0^1 \Delta f(\widetilde{X}(t-))dA(t)\right| &= \mathbb{E}\left|\int_0^1 \int_0^{\delta S(t)} f_h'(X(t-)+u)dudA(t)\right| \\
&\le \mathbb{E}\left|\int_0^1 \int_0^{\delta S(t)} \delta^{-2}\,|X(t-)+u|\,dudA(t)\right| \\
&\le \mathbb{E}\int_0^1 \delta^{-1}S(t)(X(t-)+S(t))dA(t) \\
&= \delta^{-1}(\mathbb{E}S)\mathbb{E}\int_0^1 X(t-)dA(t) + \delta^{-1}\mathbb{E}S^2\mathbb{E}A(1)
\end{aligned}$$

The second term on the right-hand side is finite because $\mathbb{E}A(1) = \lambda$ (Lemma 2). Letting $\tau_m$ be the time of the $m$th jump, the first term is bounded by

$$\mathbb{E}\int_0^1 X(t-)dA(t) \le \mathbb{E}\big(A(1)\sup_{0\le t\le 1} X(t)\big) \le \sqrt{\mathbb{E}(A(1))^2}\sqrt{\mathbb{E}\Big(X(0)+\delta\sum_{m=1}^{A(1)} S(\tau_m)\Big)^2}.$$

The right-hand side is finite because $\mathbb{E}X^2 < \infty$, $\mathbb{E}A^2(1) < \infty$ (since $\mathbb{E}U^2 < \infty$), and that $\mathbb{E}(\sum_{m=1}^{A(1)} S(\tau_m))^2 < \infty$ (since $A(1)$ is independent of $S(\tau_m)$). $\square$

## EC.4. The prelimit approach: supporting proofs

*Proof of Lemma 5* The first inequality follows from monotonicity: forcing an arrival to occur immediately can only increase the initial workload and hence can only delay the first time the workload empties.

For the second inequality, consider the system initialized with workload $x/\delta + S$ and residual interarrival time $U$. Treat the initial amount $x/\delta$ as low-priority work and all remaining work, including the initial $S$ and all future arrivals, as high-priority work. The low-priority work is processed only when no high-priority work is present, that is, during idle periods of the high-priority system. Therefore, the low-priority work is cleared once the cumulative idle time of the high-priority system exceeds $x/\delta$. If $N_x$ denotes the number of idle periods required for this to occur, then $\mathbb{E}N_x < \infty$ by Gut (1974). Since the high-priority system has i.i.d. cycles with finite mean $\mathbb{E}(\bar{B}+\bar{I})$, Wald's identity implies that the expected time until the end of the $N_x$th idle period is $\mathbb{E}N_x\,\mathbb{E}(\bar{B}+\bar{I})<\infty$. This upper-bounds $\mathbb{E}(\mathbb{E}_{x+\delta S,U}B_0)$ and proves (19). □

### EC.4.1. Proving Proposition 2

We require several auxiliary lemmas.

Lemma EC.4. *For any $h\in\mathrm{Lip}(1)$ and $M>0$,*

$$\partial_x F_h^M(z)=\mathbb{E}_z\int_0^{B_0\wedge M} h'(X(t))dt,\quad z=(x,r_a)\in\mathbb{S}.$$

*Proof of Lemma EC.4* The proof is identical to that of Lemma 6; see Appendix EC.4.2.
□

Lemma EC.5. *For any differentiable $g:\mathbb{R}_+\to\mathbb{R}$ with $\mathbb{E}|g(U)|,\mathbb{E}|g(R_a)|,\ \mathbb{E}|g'(R_a)|<\infty$,*

$$\mathbb{E}g'(R_a)=\lambda\big(\mathbb{E}g(U)-\lim_{\epsilon\to 0}g(\epsilon)\big).$$

*Proof of Lemma EC.5* Since $\mathbb{E}A(1)=\lambda$ (Lemma 2), the result follows from the BAR in Lemma 1 with $f(Z(t))=g(R_a(t))$ there. □

The next two lemmas are proved in Appendix EC.4.1.1.

Lemma EC.6. *For any $h\in\mathrm{Lip}(1)$ and almost all $M>0$,*

$$-\delta\partial_x F_h^M(z)-\partial_{r_a}F_h^M(z)=\mathbb{E}_z h(X(M))-h(x),\quad z=(x,r_a)\in\mathbb{S}.$$

*In particular, $\partial_{r_a}F_h^M(z)$ is well defined for $z\in\mathbb{S}$.*

Lemma EC.7. *For any $h\in\mathrm{Lip}(1)$ and $M>0$,*

$$\lim_{\epsilon\to 0}F_h^M(x,\epsilon)=\mathbb{E}F_h^M(x+\delta S,U).$$

*Proof of Proposition 2* For almost all $M>0$, Lemma EC.6 says that

$$-\delta\partial_x F_h^M(z)-\partial_{r_a}F_h^M(z)=\mathbb{E}_z h(X(M))-h(x),\quad z\in\mathbb{S}.\tag{EC.13}$$

Observe that $f(r_a)=F_h^M(x,r_a)$ satisfies the conditions of Lemma EC.5. Indeed, $|\mathbb{E}f(U)|,|\mathbb{E}f(R_a)|<\infty$ because $M<\infty$. Furthermore, $|\mathbb{E}f'(R_a)|<\infty$ follows from the expression for $\partial_{r_a} F_h^M(z)$ in Lemma EC.6, together with the observation that $|\partial_x F_h^M(z)| \leq M$, which follows from Lemma EC.4. Setting $r_a = R_a$ in (EC.13) and taking expected values yields

$$\begin{aligned}
\mathbb{E}\big(\mathbb{E}_{z,R_a} h(X(M)) - h(x)\big) &= -\delta \mathbb{E}\partial_x F_h^M(x,R_a) - \mathbb{E}\partial_{r_a} F_h^M(x,R_a) \\
&= -\delta \mathbb{E}\partial_x F_h^M(x,R_a) - \lambda\Big(\mathbb{E}F_h^M(x,U) - \lim_{\epsilon\to 0} F_h^M(x,\epsilon)\Big) \\
&= -\delta \mathbb{E}\partial_x F_h^M(x,R_a) - \lambda\Big(\mathbb{E}F_h^M(x,U) - \mathbb{E}F_h^M(x+\delta S,U)\Big),
\end{aligned}$$

where the second and third equalities follow from Lemmas EC.5 and EC.7, respectively. □

#### EC.4.1.1. Auxiliary lemma proofs

*Proof of Lemma EC.6* We define $\tilde{h}(x) = h(x) - \mathbb{E}h(X)$ for convenience, in which case

$$F_h^M(z) = \int_0^M \mathbb{E}_z \tilde{h}(X(t))dt, \quad z \in \mathbb{S}.$$

Our goal is to prove that

$$-\delta\partial_x F_h^M(z) - \partial_{r_a} F_h^M(z) = \mathbb{E}_z h(X(M)) - h(x), \quad z = (x, r_a) \in \mathbb{S}. \tag{EC.14}$$

Fix $z = (x, r_a) \in \mathbb{S}$ and suppose first that $x = 0$. On one hand,

$$F_h^{M+\epsilon}(0, r_a+\epsilon) = F_h^M(0, r_a+\epsilon) + \int_M^{M+\epsilon} \mathbb{E}_{0,r_a+\epsilon}\tilde{h}(X(t))dt,$$

and on the other,

$$F_h^{M+\epsilon}(0, r_a+\epsilon) = \int_0^{\epsilon} \mathbb{E}_{0,r_a+\epsilon}\tilde{h}(X(t))dt + \int_0^M \mathbb{E}_{0,r_a}\tilde{h}(X(t))dt = \epsilon\tilde{h}(0) + F_h^M(0,r_a),$$

Equating the two expressions and dividing both sides by $\epsilon$ yields

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\big(F_h^M(0,r_a+\epsilon) - F_h^M(0,r_a)\big) = \tilde{h}(0) - \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_M^{M+\epsilon}\mathbb{E}_{0,r_a+\epsilon}\tilde{h}(X(t))dt.$$

The left-hand side equals $\delta\partial_x F_h^M(z) + \partial_{r_a} F_h^M(z) = \partial_{r_a} F_h^M(z)$, since $\partial_x F_h^M(0,r_a) = 0$ by Lemma EC.4. Thus, to prove (EC.14) when $x = 0$, it suffices to show that

$$\begin{aligned}
&\frac{1}{\epsilon}\int_M^{M+\epsilon}\big(\mathbb{E}_{0,r_a+\epsilon}\tilde{h}(X(t)) - \mathbb{E}_{0,r_a}\tilde{h}(X(t))\big)dt \\
&= \frac{1}{\epsilon}\int_M^{M+\epsilon}\mathbb{E}_{0,r_a}\big(\tilde{h}(X(t-\epsilon)) - \tilde{h}(X(t))\big)dt \to 0
\end{aligned} \tag{EC.15}$$

as $\epsilon \to 0$, which implies that

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_M^{M+\epsilon}\mathbb{E}_{0,r_a+\epsilon}\tilde{h}(X(t))dt = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_M^{M+\epsilon}\mathbb{E}_{0,r_a}\tilde{h}(X(t))dt = \mathbb{E}_{0,r_a}\tilde{h}(X(M)).$$

Observe that $|X(t-\epsilon)-X(t)|$ is bounded by the workload processed during $[t-\epsilon,t]$, which is at most $\delta\epsilon$, plus any new work that arrives during $[t-\epsilon,t]$. Letting $A([t_1,t_2])$ denote the number of customers arriving during $[t_1,t_2]$, Wald's identity says that the expected workload to arrive during $[t_1,t_2]$ equals $\mathbb{E}S\mathbb{E}A([t_1,t_2])$. Thus, to prove (EC.15), we observe that for any $h\in \mathrm{Lip}(1)$ and for all $t\in[M,M+\epsilon]$,

$$\begin{aligned}
\mathbb{E}_{0,r_a}\left|\tilde{h}(X(t-\epsilon))-\tilde{h}(X(t))\right| &\le \mathbb{E}_{0,r_a}\left|X(t-\epsilon)-X(t)\right| \\
&\le \delta\epsilon + \mathbb{E}_{0,r_a}\big(\delta\mathbb{E}S\mathbb{E}(A([t-\epsilon,t]))\big) \\
&\le \delta\epsilon + \delta\mathbb{E}S\mathbb{E}_{0,r_a}\big(A([M-\epsilon,M+\epsilon])\big).
\end{aligned}$$

It suffices to argue that the right-hand side goes to zero as $\epsilon\to 0$. By the dominated convergence theorem,

$$\lim_{\epsilon\to 0}\mathbb{E}_{0,r_a}\big(A([M-\epsilon,M+\epsilon])\big) = \mathbb{E}_{0,r_a}\big(A([M,M])\big),$$

which equals the expected number of arrivals at time $M$. The right-hand side may be non-zero if the distribution of $U$ has point masses. However, since the number of point masses is at most countable, then $\mathbb{E}_{0,r_a}\big(A([M,M])\big)=0$ for all but at most countably many $M$. This proves (EC.14) when $x=0$.

The case when $x>0$ follows similarly. We repeat the arguments, highlighting the differences. Given $z=(x,r_a)$, fix $\epsilon < x/\delta$. Then

$$F_h^{M+\epsilon}(x,r_a+\epsilon) = F_h^M(x,r_a+\epsilon) + \int_M^{M+\epsilon}\mathbb{E}_{x,r_a+\epsilon}\tilde{h}(X(t))dt$$

and

$$F_h^{M+\epsilon}(x,r_a+\epsilon) = \int_0^{\epsilon}\mathbb{E}_{x,r_a+\epsilon}\tilde{h}(X(t))dt + F_h^M(x-\delta\epsilon,r_a).$$

Equating both expressions, subtracting $F_h^M(x,r_a)$ from each side, and dividing by $\epsilon$ yields

$$\begin{aligned}&\frac{1}{\epsilon}\big(F_h^M(x,r_a+\epsilon)-F_h^M(x,r_a)\big)\\&=\frac{1}{\epsilon}\big(F_h^M(x-\delta\epsilon,r_a)-F_h^M(x,r_a)\big)+\frac{1}{\epsilon}\int_0^{\epsilon}\mathbb{E}_{x,r_a+\epsilon}\tilde{h}(X(t))dt-\frac{1}{\epsilon}\int_M^{M+\epsilon}\mathbb{E}_{x,r_a+\epsilon}\tilde{h}(X(t))dt.\end{aligned}$$

We now argue that each of the terms on the right-hand side has a well-defined limit as $\epsilon\to 0$, implying that the left-hand side converges to $\partial_x F_h^M(z)$, which is itself well defined. The first term on the right-hand side converges to $-\partial_x F_h^M(z)$, which we know exists for all $z\in\mathbb{S}$ by Lemma EC.4. Furthermore,

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_0^{\epsilon}\mathbb{E}_{x,r_a+\epsilon}\tilde{h}(X(t))dt=\tilde{h}(x)\quad\text{ and }\quad\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_M^{M+\epsilon}\mathbb{E}_{x,r_a+\epsilon}\tilde{h}(X(t))dt=\mathbb{E}_{x,r_a}\tilde{h}(X(M)).$$

The first equality is straightforward because no arrival occurs during $[0,\epsilon]$, while the second equality is proved the same way as (EC.15). □

*Proof of Lemma EC.7* Define $\tilde{h}(x)=h(x)-\mathbb{E}h(X)$ and consider first the case when $x=0$. Then

$$\begin{aligned}F_h^M(0,\epsilon)&=\int_0^{\epsilon}\mathbb{E}_{0,\epsilon}\tilde{h}(X(t))dt+\int_{\epsilon}^{M}\mathbb{E}_{0,\epsilon}\tilde{h}(X(t))dt\\&=\int_0^{\epsilon}\mathbb{E}_{0,\epsilon}\tilde{h}(X(t))dt+\mathbb{E}\int_0^{M-\epsilon}\mathbb{E}_{\delta S,U}\tilde{h}(X(t))dt,\end{aligned}$$

where the outer expectation is with respect to $U$ and $S$. Taking $\epsilon\to 0$, the left-hand side converges to $\lim_{\epsilon\to 0}F_h^M(0,\epsilon)$ while the right-hand side converges to

$$\lim_{\epsilon\to 0}\mathbb{E}\int_0^{M-\epsilon}\mathbb{E}_{\delta S,U}\tilde{h}(X(t))dt=\mathbb{E}\int_0^{M}\mathbb{E}_{\delta S,U}\tilde{h}(X(t))dt-\lim_{\epsilon\to 0}\mathbb{E}\int_{M-\epsilon}^{M}\mathbb{E}_{\delta S,U}\tilde{h}(X(t))dt.$$

The first term equals $\mathbb{E}F_h^M(\delta S,U)$ while the second term is zero because $h\in\mathrm{Lip}(1)$. Now suppose that $x>0$ and take $\epsilon<x/\delta$. Arguing as before,

$$F_h^M(x,\epsilon)=\int_0^{\epsilon}\mathbb{E}_{x,\epsilon}\tilde{h}(X(t))dt+\mathbb{E}F_h^{M-\epsilon}(x-\delta\epsilon+\delta S,U).$$

To conclude, we use the fundamental theorem of calculus to write

$$\mathbb{E}F_h^{M-\epsilon}(x-\delta\epsilon+\delta S,U)=\mathbb{E}F_h^{M-\epsilon}(x+\delta S,U)+\mathbb{E}\int_0^{-\delta\epsilon}\partial_x F_h^{M-\epsilon}(x+v+\delta S,U)dv.$$

The second term on the right-hand side converges to zero because $\left|\partial_x F_h^{M-\epsilon}(z)\right| \leq M$ due to Lemma EC.4. The first term converges to $\mathbb{E}F_h^M(x+\delta S, U)$ because

$$\lim_{\epsilon\to 0}\mathbb{E}\int_0^{M-\epsilon}\mathbb{E}_{x+\delta S,U}\tilde{h}(X(t))dt = \mathbb{E}\int_0^M \mathbb{E}_{x+\delta S,U}\tilde{h}(X(t))dt - \lim_{\epsilon\to 0}\mathbb{E}\int_{M-\epsilon}^M \mathbb{E}_{x+\delta S,U}\tilde{h}(X(t))dt,$$

and the second term equals zero since $h\in \mathrm{Lip}(1)$. □

### EC.4.2. Proofs of Lemmas 6 and 7

We recall the synchronous coupling $\{Z^{(\epsilon)}(t): t\geq 0\}$ defined in Section 3.3.

*Proof of Lemma 6* First, observe that

$$\frac{1}{\epsilon}\mathbb{E}\Big(\int_0^\infty \big(\mathbb{E}_{x+\epsilon,T}h(X(t)) - \mathbb{E}_{x,T}h(X(t))\big)dt\Big)$$
$$= \frac{1}{\epsilon}\mathbb{E}\Big(\mathbb{E}_{x,T}\int_0^{B_0}\big(h(X^{(\epsilon)}(t)) - h(X(t))\big)dt\Big) + \frac{1}{\epsilon}\mathbb{E}\Big(\mathbb{E}_{x,T}\int_{B_0}^{B_0^{(\epsilon)}}\big(h(X^{(\epsilon)}(t)) - h(X(t))\big)dt\Big).$$

Note that $\left|h(X^{(\epsilon)}(t)) - h(X(t))\right|/\epsilon \leq \|h'\| \leq 1$ and $B_0^{(\epsilon)} \to B_0$ as $\epsilon\to 0$. Also note that for all $\epsilon < 1$,

$$\mathbb{E}_{x,T}B_0^{(\epsilon)} = \mathbb{E}_{x+\epsilon,T}B_0 \leq \mathbb{E}_{x+1,T}B_0 \leq \mathbb{E}\big(\mathbb{E}_{x+1+\delta S,U}B_0\big) < \infty,$$

where the second-last inequality follows from the fact that the busy period starting at state $(x+1,T)$ is made longer if the next arrival happens immediately. The DCT then implies that

$$\frac{1}{\epsilon}\mathbb{E}\Big(\mathbb{E}_{x,T}\int_0^{B_0}\big(h(X^{(\epsilon)}(t)) - h(X(t))\big)dt\Big) \to \mathbb{E}\Big(\mathbb{E}_{x,T}\int_0^{B_0} h'(X(t))dt\Big),$$
$$\frac{1}{\epsilon}\mathbb{E}\Big(\mathbb{E}_{x,T}\int_{B_0}^{B_0^{(\epsilon)}}\big(h(X^{(\epsilon)}(t)) - h(X(t))\big)dt\Big) \to 0.$$

□

*Proof of Lemma 7* Fix $h\in\mathrm{Lip}(1)$. Let $\mu_{R_a}$ and $\mu_{X|r_a}$ denote the law of $R_a$ and conditional law of $X$ given $R_a = r_a$, respectively. We first prove (30). Note that

$$\mathbb{E}h(X) = \int_0^\infty\int_0^\infty \mathbb{E}_{y,r_a}h(X(M))d\mu_{X|r_a}(y)d\mu_{R_a}(r_a),$$
$$\mathbb{E}\big(\mathbb{E}_{x,R_a}h(X(M))\big) = \int_0^\infty \mathbb{E}_{x,r_a}h(X(M))d\mu_{R_a}(r_a) = \int_0^\infty\int_0^\infty \mathbb{E}_{x,r_a}h(X(M))d\mu_{X|r_a}(y)d\mu_{R_a}(r_a).$$

The first equality is true because $\mathbb{E}h(X)$ coincides with $\mathbb{E}h(X(M))$ if $Z(0)$ is initialized according to $Z$. It follows that

$$\mathbb{E}\big(\mathbb{E}_{x,R_a}h(X(M))\big) - \mathbb{E}h(X) = \int_0^\infty \int_0^\infty \big(\mathbb{E}_{x,r_a}h(X(M)) - \mathbb{E}_{y,r_a}h(X(M))\big)d\mu_{X|r_a}(y)d\mu_{R_a}(r_a).$$

Since $h \in \mathrm{Lip}(1)$, we can use the synchronous coupling defined in (25) to compare two workload processes initialized at $(x, r_a)$ and $(y, r_a)$. Under this coupling, their workloads differ by at most $|x - y|$ until they couple. Coupling occurs no later than the first time the larger initial workload process empties. Therefore,

$$|\mathbb{E}_{x,r_a}h(X(M)) - \mathbb{E}_{y,r_a}h(X(M))| \leq |x - y|\, \mathbb{P}_{x\vee y,r_a}\big(B_0 > M\big),$$

where the probability on the right-hand side is the probability that the coupled processes have not yet coupled by time $M$. Thus,

$$\lim_{M\to\infty} \big|\mathbb{E}\big(\mathbb{E}_{x,R_a}h(X(M))\big) - \mathbb{E}h(X)\big| \leq \lim_{M\to\infty} \mathbb{E}\Big(|x - X|\, \mathbb{P}_{x\vee X,R_a}\big(B_0 > M\big)\Big) = 0,$$

where the expectation on the right-hand side is taken with respect to the joint law of $(X, R_a)$. The last equality follows from the DCT because $\mathbb{E}X < \infty$, and because $\lim_{M\to\infty} \mathbb{P}_{x\vee x',r_a}\big(B_0 > M\big) = 0$ for any $x, x', r_a > 0$ by (19). To prove (31), one can reuse the arguments used to prove Lemma 6 to show that

$$\partial_x F_h^M(x, r_a) = \mathbb{E}_{x,r_a} \int_0^{B_0\wedge M} h'(X(t))dt \to \partial_x F_h(x, r_a) \qquad \text{as } M \to \infty \text{ for all } (x, r_a) \in \mathbb{S},$$

and also that $\lim_{M\to\infty} \mathbb{E}\partial_x F_h^M(x, R_a) = \mathbb{E}\lim_{M\to\infty} \partial_x F_h^M(x, R_a)$. Lastly, we prove (32). Similar to the way we argued (26),

$$\begin{aligned} F_h^M(x+\epsilon, r_a) - F_h^M(x, r_a) &= \mathbb{E}_{x,r_a} \int_0^{B_0^{(\epsilon)}\wedge M} \big(h(X^{(\epsilon)}(t)) - h(X(t))\big)dt \\ &\to F_h(x+\epsilon, r_a) - F_h(x, r_a), \qquad \text{as } M \to \infty \text{ for all } (x, r_a) \in \mathbb{S}. \end{aligned}$$

Let $\hat{h}(x) = x$ and observe that since $h \in \mathrm{Lip}(1)$ and $X^{(\epsilon)}(t) \geq X(t)$, then

$$\big|F_h^M(x+\epsilon, r_a) - F_h^M(x, r_a)\big| \leq \mathbb{E}_{x,r_a} \int_0^{B_0^{(\epsilon)}\wedge M} \big(X^{(\epsilon)}(t) - X(t)\big)dt \leq F_{\hat{h}}(x+\epsilon, r_a) - F_{\hat{h}}(x, r_a).$$

It remains to show that $\mathbb{E}\big(F_{\hat{h}}(x+\delta S,U)-F_{\hat{h}}(x,U)\big)<\infty$, because then we can use the DCT to conclude (32). The finiteness of this expectation follows from

$$\begin{aligned}\mathbb{E}\big(F_{\hat{h}}(x+\delta S,U)-F_{\hat{h}}(x,U)\big)&=\lim_{M\to\infty}\mathbb{E}\big(F_{\hat{h}}^M(x+\delta S,U)-F_{\hat{h}}^M(x,U)\big)\\&=\lim_{M\to\infty}\mathbb{E}\big(\mathbb{E}_{x,R_a}X(M)-x\big)+\lim_{M\to\infty}\delta\mathbb{E}\partial_x F_{\hat{h}}^M(x,R_a),\end{aligned}$$

where the first equality is due to the monotone convergence theorem, since $F_{\hat{h}}^M(x+\epsilon,r_a)-F_{\hat{h}}^M(x,r_a)$ is increasing in $M$ and is nonnegative for any $(x,r_a)\in\mathbb{S}$, and the second equality is due to (23). The right-hand side is finite by (30) and (31). □

### EC.4.3. Section 3.4 Proofs

*Proof of Lemma 8* Recall that $J(x,r_a)=-(x\wedge\delta r_a)+\delta S'$. It follows that

$$\begin{aligned}F_h(x+\delta s,r_a)-F_h(x,r_a)&=\int_0^\infty\big(\mathbb{E}_{x+\delta s,r_a}h(X(t))-\mathbb{E}_{x,r_a}h(X(t))\big)dt\\&=\int_0^{r_a}\Big(h\big((x+\delta s-\delta t)^+\big)-h\big((x-\delta t)^+\big)\Big)dt\\&\quad+\mathbb{E}\Big(F_h\big(x+\delta s+J(x+\delta s,r_a),U\big)-F_h\big(x+J(x,r_a),U\big)\Big).\end{aligned}$$

To conclude, note that

$$\begin{aligned}&\mathbb{E}\Big(F_h\big(x+\delta s+J(x+\delta s,r_a),U\big)-F_h\big(x+J(x,r_a),U\big)\Big)\\&=\mathbb{E}\Big(F_h\big(x+\delta s+J(x,r_a),U\big)-F_h\big(x+J(x,r_a),U\big)\Big)\\&\quad+\mathbb{E}\Big(F_h\big(x+\delta s+J(x+\delta s,r_a),U\big)-F_h\big(x+\delta s+J(x,r_a),U\big)\Big).\end{aligned}$$

Using the fundamental theorem of calculus, together with Lemma 6, which shows that $\partial_x\mathbb{E}F_h(x+\delta S,U)=\mathbb{E}\partial_x F_h(x+\delta S,U)$, we arrive at

$$\begin{aligned}&\mathbb{E}\Big(F_h\big(x+\delta s+J(x+\delta s,r_a),U\big)-F_h\big(x+\delta s+J(x,r_a),U\big)\Big)\\&=\mathbb{E}\Big(\int_{-x\wedge(\delta r_a)}^{-(x+\delta s)\wedge(\delta r_a)}\partial_x F_h(x+\delta s+v+\delta S',U)dv\Big)\\&=\mathbb{E}^{S'}\Big(\int_{-x\wedge(\delta r_a)}^{-(x+\delta s)\wedge(\delta r_a)}\mathbb{E}^U\partial_x F_h(x+\delta s+v+\delta S',U)dv\Big)\\&=\mathbb{E}^{S'}\Big(\int_{-x\wedge(\delta r_a)}^{-(x+\delta s)\wedge(\delta r_a)}\partial_x\mathbb{E}^U F_h(x+\delta s+v+\delta S',U)dv\Big).\end{aligned}$$

Interchanging $\mathbb{E}^U$ with the integral in the second equality is justified by the Fubini-Tonelli theorem because $\mathbb{E}^{S'}\mathbb{E}^U\left|\partial_x F_h(x+\delta S',U)\right| \le \mathbb{E}^{S'}\mathbb{E}^U\mathbb{E}_{x+\delta S',U}B_0 < \infty$ for all $x\ge 0$ by Lemma 6 and (19).

$\square$

**EC.4.3.1. Proving Lemma 9** We recall that $\bar F_h'(x) = \partial_x \mathbb{E}F_h(x,R_a)$ and that $\bar F_h''(x)$ and $\bar F_h'''(x)$ are assumed to exist. We recall (35), or

$$\begin{aligned}&\mathbb{E}h(X) - \mathbb{E}h(x+J(x,R_a'))\\ =\ & -\delta\mathbb{E}\bar F_h'(x+J(x,R_a')) + \lambda\mathbb{E}\big(F_h(x+\delta S,R_a') - F_h(x,R_a')\big) - \mathbb{E}\big(\epsilon(x,R_a',S)\big),\end{aligned} \tag{EC.16}$$

where $J(x,r_a) = -(x\wedge \delta r_a) + \delta S'$. The following lemma expands the first two terms on the right-hand side of (EC.16). We prove it after proving Lemma 9.

Lemma EC.8. *For any $x\ge 0$,*

$$\begin{aligned}\bar F_h'(x+J(x,R_a')) = \bar F_h'(x) + \delta(S'-R_a')\bar F_h''(x) + 1(\delta R_a' < x)\int_0^{\delta(S'-R_a')}\int_0^v \bar F_h'''(x+u)dudv\\ + 1(\delta R_a'\ge x)\big(\bar F_h'(\delta S') - \bar F_h'(x) - \delta(S'-R_a')\bar F_h''(x)\big)\end{aligned}$$

$$\mathbb{E}\big(F_h(x+\delta S,R_a') - F_h(x,R_a')\big) = \delta\mathbb{E}S\bar F_h'(x) + \frac{1}{2}\delta^2\mathbb{E}S^2\bar F_h''(x) + \mathbb{E}\int_0^{\delta S}(\delta S - v)\int_0^v \bar F_h'''(x+u)dudv,$$

*Proof of Lemma 9* Recall that $\lambda\mathbb{E}S = \rho$. Combining Lemma EC.8 with (EC.16) yields

$$\begin{aligned}&\mathbb{E}h(X) - \mathbb{E}h(x+J(x,R_a'))\\ =\ & -\delta\big(\bar F_h'(x) + \delta\mathbb{E}(S'-R_a')\bar F_h''(x)\big) + \lambda\big(\delta\mathbb{E}S\bar F_h'(x) + \frac{1}{2}\delta^2\mathbb{E}S^2\bar F_h''(x)\big)\\ & -\delta\mathbb{E}\Big(1(\delta R_a'\ge x)\big(\bar F_h'(\delta S') - \bar F_h'(x) - \delta(S'-R_a')\bar F_h''(x)\big)\Big)\\ & -\delta\mathbb{E}\Big(1(\delta R_a' < x)\int_0^{\delta(S'-R_a')}\int_0^v \bar F_h'''(x+u)dudv\Big)\\ & +\lambda\mathbb{E}\int_0^{\delta S}(\delta S - v)\int_0^v \bar F_h'''(x+u)dudv.\end{aligned}$$

Using the facts that $\lambda\mathbb{E}S=\rho$, $\lambda\mathbb{E}U = 1$, and that $\mathbb{E}R_a' = \lambda\mathbb{E}U^2/2$, we see that the first line on the right-hand side equals

$$-\delta(1-\rho)\bar F_h'(x) + \frac{1}{2}\delta^2\big(\lambda\mathbb{E}S^2 - 2\lambda\mathbb{E}U\mathbb{E}S' + \lambda\mathbb{E}U^2\big)\bar F_h''(x).$$

Let us call this term $G_{Y_2}\bar{F}_h(x)$. Since $\bar{F}_h'(0)=0$ due to Lemma 6 (because $B_0=0$ if the initial workload $X(0)=0$), our assumptions that $\mathbb{E}\left|\bar{F}_h'(Y_2)\right|, \mathbb{E}\left|\bar{F}_h''(Y_2)\right| < \infty$ and integration by parts yield $\mathbb{E}G_{Y_2}\bar{F}_h(Y_2)=0$. □

*Proof of Lemma EC.8* The expression for $\bar{F}_h'(x+J(x,R_a'))$ follows from the facts that

$$\bar{F}_h'(x+J(x,R_a')) = 1(\delta R_a' \geq x)\bar{F}_h'(\delta S') + 1(\delta R_a' < x)\bar{F}_h'(x-\delta R_a'+\delta S')$$

and, for all $x>\delta R_a'$,

$$\bar{F}_h'(x-\delta R_a'+\delta S') = \bar{F}_h'(x) + \delta(S'-R_a')\bar{F}_h''(x) + \int_0^{\delta(S'-R_a')}\int_0^v \bar{F}_h'''(x+u)dudv.$$

Next, we argue that

$$\mathbb{E}\big(F_h(x+\delta S, R_a') - F_h(x,R_a')\big) = \mathbb{E}\int_0^{\delta S}\bar{F}_h'(x+v)dv, \tag{EC.17}$$

so that the expression for $\mathbb{E}\big(F_h(x+\delta S,R_a')-F_h(x,R_a')\big)$ also follows from Taylor expansion of the integrand around $x$. To prove (EC.17), note that

$$\begin{aligned}
\mathbb{E}\big(F_h(x+\delta S, R_a') - F_h(x,R_a')\big) &= \mathbb{E}\int_0^{\delta S}\partial_x F_h(x+v,R_a')dv \\
&= \mathbb{E}^S\int_0^{\delta S}\mathbb{E}^{R_a'}\partial_x F_h(x+v,R_a')dv \\
&= \mathbb{E}^S\int_0^{\delta S}\partial_x \mathbb{E}F_h(x+v,R_a')dv \\
&= \mathbb{E}\int_0^{\delta S}\bar{F}_h'(x+v)dv.
\end{aligned}$$

The first and second-last equalities follows from Lemma 6. Once we justify the interchange of the integral and expectation in the second equality using the Fubini-Tonelli theorem, (EC.17) will follow. Let $\hat{h}(x)=x$. Using the form of $\partial_x F_h(x,r_a)$ from Lemma 6, it follows that for any $h\in \mathrm{Lip}(1)$,

$$|\partial_x F_h(x,r_a)| \leq \mathbb{E}_{x,r_a}B_0 = \mathbb{E}_{x,r_a}\int_0^{B_0}\hat{h}'(X(t))dt = \partial_x F_{\hat{h}}(x,r_a).$$

Thus,

$$\mathbb{E}\int_0^{\delta S}|\partial_x F_h(x+v,R_a')|\,dv \leq \mathbb{E}\int_0^{\delta S}\partial_x F_{\hat{h}}(x+v,R_a')dv = \mathbb{E}\big(F_{\hat{h}}(x+\delta S,R_a') - F_{\hat{h}}(x,R_a')\big),$$

and the right-hand side is finite because the right-hand side of (32) in Lemma 7 is finite.
□

## EC.5. $G/G/1$ workload Stein factor bounds: supporting proofs

For the entirety of this section we fix $h(x) = x$ and assume that $\overline{\eta} = \sup\{\eta(x) : x \geq 0\} < \infty$.

### EC.5.1. Second-order bounds (Proof of Lemma 14)

Recall that $G(x) = \mathbb{P}(U \leq x)$. The following auxiliary lemma is needed to prove Lemma 14.

LEMMA EC.9. *For any $\epsilon > 0$ and $(x, r_a) \in \mathbb{S}$ with $r_a < x/\delta$,*

$$\frac{1}{\epsilon}\mathbb{P}_{x,r_a}(R_a(B_0) < \epsilon/\delta) \leq \overline{\eta}/\delta, \tag{EC.18}$$

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{P}_{x,r_a}(R_a(B_0) < \epsilon/\delta) = \frac{1}{\delta}\mathbb{E}_{x,r_a}\eta(\alpha(B_0)). \tag{EC.19}$$

*Proof of Lemma EC.9* Let $U_n$ denote the interarrival time of the $n$th customer, let $W_0 = V(0)$, and let $W_n = V(U_1 + \cdots + U_n)$ be the workload in the system right after the $n$th customer arrives, which includes the workload brought by the $n$th customer. Let

$$\sigma = \min\{n \geq 1 : U_n > W_{n-1}\}$$

be the number of customers served in the first busy period $[0, B_0]$. Now assuming that $Z(0) = (x, r_a) \in \mathbb{S}$ with $r_a < x/\delta$, it must be that $\sigma > 1$, because $W_0 = x/\delta$ and $U_1 = r_a$. Since $\{R_a(B_0) \leq \epsilon/\delta\} = \{U_\sigma \leq W_{\sigma-1} + \epsilon/\delta\}$, it follows that

$$\begin{aligned}
\frac{1}{\epsilon}\mathbb{P}_{x,r_a}(R_a(B_0) \leq \epsilon/\delta) &= \frac{1}{\epsilon}\sum_{n=2}^{\infty}\mathbb{P}_{x,r_a}\big(U_n \leq W_{n-1} + \epsilon/\delta \big| \sigma = n\big)\mathbb{P}_{x,r_a}(\sigma = n) \\
&= \frac{1}{\epsilon}\sum_{n=2}^{\infty}\mathbb{E}_{x,r_a}\Big[\mathbb{P}_{x,r_a}\big(U_n \leq W_{n-1} + \epsilon/\delta \big| \sigma = n, W_{n-1}\big)\Big| \sigma = n\Big]\mathbb{P}_{x,r_a}(\sigma = n).
\end{aligned}$$

To proceed, note that $\{\sigma = n\} = \{U_1 \leq W_0, \ldots, U_{n-1} \leq W_{n-2}, U_n > W_{n-1}\}$ for any $n \geq 1$, implying that for any $n \geq 2$,

$$\begin{aligned}
&\mathbb{P}_{x,r_a}\big(U_n \leq W_{n-1} + \epsilon/\delta \big| \sigma = n, W_{n-1}\big) \\
&= \mathbb{P}\big(U_n \leq W_{n-1} + \epsilon/\delta \big| W_0 = x/\delta, U_1 = r_a, \sigma = n, W_{n-1}\big) \\
&= \mathbb{P}\big(U_n \leq W_{n-1} + \epsilon/\delta \big| W_0 = x/\delta, U_1 = r_a, U_1 \leq W_0, \ldots, U_{n-1} \leq W_{n-2}, U_n > W_{n-1}, W_{n-1}\big) \\
&= \mathbb{P}\big(U_n \leq W_{n-1} + \epsilon/\delta \big| U_n > W_{n-1}, W_{n-1}\big) \\
&= \mathbb{P}\big(U \leq W_{n-1} + \epsilon/\delta \big| U > W_{n-1}, W_{n-1}\big) \\
&= \frac{G(W_{n-1} + \epsilon/\delta) - G(W_{n-1})}{1 - G(W_{n-1})},
\end{aligned}$$

and therefore

$$\frac{1}{\epsilon}\mathbb{P}_{x,r_a}(R_a(B_0)\le\epsilon/\delta)=\frac{1}{\epsilon}\sum_{n=1}^{\infty}\mathbb{E}_{x,r_a}\left[\frac{G(W_{n-1}+\epsilon/\delta)-G(W_{n-1})}{1-G(W_{n-1})}\middle|\sigma=n\right]\mathbb{P}_{x,r_a}(\sigma=n)$$
$$=\frac{1}{\epsilon}\mathbb{E}_{x,r_a}\left[\frac{G(W_{\sigma-1}+\epsilon/\delta)-G(W_{\sigma-1})}{1-G(W_{\sigma-1})}\right]. \tag{EC.20}$$

To prove (EC.18), observe that the right-hand side of (EC.20) is bounded by $\overline{\eta}/\delta$ because by the mean value theorem,

$$\frac{G(w+\epsilon/\delta)-G(w)}{1-G(w)}=\frac{\epsilon}{\delta}\frac{G'(\xi)}{1-G(w)}=\frac{\epsilon}{\delta}\eta(\xi)\frac{1-G(\xi)}{1-G(w)}\le\frac{\epsilon}{\delta}\overline{\eta}$$

for some $\xi\in[w,w+\epsilon/\delta]$, where the last inequality follows from $\xi\ge w$ and $\eta(x)\le\overline{\eta}$. Once we observe that $W_{\sigma-1}=\alpha(B_0)$, then (EC.19) follows from taking $\epsilon\to0$ in (EC.20) and applying the dominated convergence theorem. □

*Proof of Lemma 14* Fix $h\in\mathcal{M}_2$, $x\ge0$, and $\epsilon>0$, and consider

$$\frac{1}{\epsilon}\big(\partial_x\mathbb{E}F_h(x+\epsilon,T)-\partial_x\mathbb{E}F_h(x,T)\big)$$
$$=\frac{1}{\epsilon}\mathbb{E}\left(\mathbb{E}_{x,T}\int_0^{B_0}\big(h'(X^{(\epsilon)}(t))-h'(X(t))\big)dt\right)+\frac{1}{\epsilon}\mathbb{E}\left(\mathbb{E}_{x,T}\int_{B_0}^{B_0^{(\epsilon)}}h'(X^{(\epsilon)}(t))dt\right).$$

Repeating the proof of Lemma 6 yields

$$\frac{1}{\epsilon}\mathbb{E}\left(\mathbb{E}_{x,T}\int_0^{B_0}\big(h'(X^{(\epsilon)}(t))-h'(X(t))\big)dt\right)\to\partial_x\mathbb{E}F_{h'}(x,T)\qquad\text{as }\epsilon\to0.$$

Recall that $R_a(B_0)$ is the residual interarrival time at the end of the initial busy period (which also equals the length of the first idle period $I_0$). If $R_a(B_0)\ge\epsilon/\delta$, then there is no arrival during the interval $[B_0,B_0^{(\epsilon)})$. Since $X^{(\epsilon)}(B_0)=\epsilon$, this implies that

$$\frac{1}{\epsilon}\mathbb{E}\left(\mathbb{E}_{x,T}\left(1(R_a(B_0)\ge\epsilon/\delta)\int_{B_0}^{B_0^{(\epsilon)}}h'(X^{(\epsilon)}(t))dt\right)\right)$$
$$=\mathbb{E}\Big(\mathbb{P}_{x,T}\big(R_a(B_0)\ge\epsilon/\delta\big)\Big)\frac{1}{\epsilon}\int_0^{\epsilon/\delta}h'(\epsilon-\delta t)dt.$$

As $\epsilon\to0$, the right-hand side converges to

$$\mathbb{E}\Big(\mathbb{P}_{x,T}\big(R_a(B_0)>0\big)\Big)\frac{1}{\delta}h'(0)=\frac{1}{\delta}h'(0).$$

To justify the last equality, we observe that $R_a(B_0)=0$ would imply that an arrival occurs precisely at the instant that the workload hits zero. Since the workload process is right-continuous, this would imply that $X(B_0)>0$, which contradicts the definition of $B_0$. It remains to show that

$$\begin{aligned}&\frac{1}{\epsilon}\mathbb{E}\bigg(\mathbb{E}_{x,T}\Big(1(R_a(B_0)<\epsilon/\delta)\int_{B_0}^{B_0^{(\epsilon)}} h'(X^{(\epsilon)}(t))dt\Big)\bigg)\\ &\to \frac{1}{\delta}\Big(\theta(x/\delta)+\mathbb{E}\big(1(T<x/\delta)\mathbb{E}_{x,T}\eta\big(\alpha(B_0)\big)\big)\Big)\mathbb{E}\big(\partial_x F_h(\delta S,U)\big).\end{aligned}$$

Since $R_a(B_0)<\epsilon/\delta$ implies that an arrival occurs in $[B_0,B_0^{(\epsilon)})$, then

$$\begin{aligned}&\frac{1}{\epsilon}\mathbb{E}\bigg(\mathbb{E}_{x,T}\Big(1(R_a(B_0)<\epsilon/\delta)\int_{B_0}^{B_0^{(\epsilon)}} h'(X^{(\epsilon)}(t))dt\Big)\bigg)\\ &=\frac{1}{\epsilon}\mathbb{E}\bigg(\mathbb{E}_{x,T}\Big(1(R_a(B_0)<\epsilon/\delta)\int_{0}^{R_a(B_0)} h'(\epsilon-\delta t)dt\Big)\bigg)\\ &\quad+\frac{1}{\epsilon}\mathbb{E}\bigg(\mathbb{E}_{x,T}\Big(1(R_a(B_0)<\epsilon/\delta)\mathbb{E}\big(\partial_x F_h(\epsilon-\delta R_a(B_0)+\delta S,U)\big)\Big)\bigg).\end{aligned}$$

The first term on the right-hand side converges to zero as $\epsilon\to 0$. To analyze the second term, note that $R_a(B_0)<\epsilon/\delta$ implies that $T<x/\delta+\epsilon/\delta$, and if $x/\delta\le T<x/\delta+\epsilon/\delta$ then $R_a(B_0)=T-x/\delta$. Therefore, the second term equals

$$\begin{aligned}&\frac{1}{\epsilon}\mathbb{E}\bigg(1(T<x/\delta)\mathbb{E}_{x,T}\Big(1(R_a(B_0)<\epsilon/\delta)\mathbb{E}\big(\partial_x F_h(\epsilon-\delta R_a(B_0)+\delta S,U)\big)\Big)\bigg)\\ &+\frac{1}{\epsilon}\mathbb{E}\bigg(1(x/\delta\le T<x/\delta+\epsilon/\delta)\mathbb{E}_{x,T}\Big(\mathbb{E}\big(\partial_x F_h(\epsilon-(\delta T-x)+\delta S,U)\big)\Big)\bigg).\end{aligned}$$

It is straightforward to check that $\sup_{0\le x'\le\epsilon}\left|\mathbb{E}\big(\partial_x F_h(x'+\delta S,U)\big)-\mathbb{E}\big(\partial_x F_h(\delta S,U)\big)\right|\to 0$ as $\epsilon\to 0$. Lemma EC.9 and the DCT then yield

$$\begin{aligned}&\frac{1}{\epsilon}\mathbb{E}\bigg(1(T<x/\delta)\mathbb{E}_{x,T}\Big(1(R_a(B_0)<\epsilon/\delta)\mathbb{E}\big(\partial_x F_h(\epsilon-\delta R_a(B_0)+\delta S,U)\big)\Big)\bigg)\\ &\to\frac{1}{\delta}\mathbb{E}\big(1(T<x/\delta)\mathbb{E}_{x,T}\eta\big(\alpha(B_0)\big)\big)\mathbb{E}\big(\partial_x F_h(\delta S,U)\big).\end{aligned}$$

Similarly, using the fact that $\theta(x)$ is bounded,

$$\begin{aligned}&\frac{1}{\epsilon}\mathbb{E}\bigg(1(x/\delta\le T<x/\delta+\epsilon/\delta)\mathbb{E}_{x,T}\Big(\mathbb{E}\big(\partial_x F_h(\epsilon-(\delta T-x)+\delta S,U)\big)\Big)\bigg)\\ &\to\frac{1}{\delta}\theta(x/\delta)\mathbb{E}\big(\partial_x F_h(\delta S,U)\big).\end{aligned}$$

□

### EC.5.2. Third-order bounds

**EC.5.2.1. The renewal process driven by idle times** In this section we prove Lemma 16. Recall that $\alpha(t)$ is the age of the interarrival process at time $t\geq 0$. Further recall from the discussion following (17) that $B_n$ and $I_n$, $n\geq 0$, are the lengths of the $n$th busy and idle periods, respectively. While $B_0$ and $I_0$ are the initial busy and idle period lengths, which depend on the initial condition $Z(0)$, the pairs $(B_n, I_n)$, $n\geq 1$, are i.i.d. with distributions $\bar{B}$ and $\bar{I}$, respectively. Given $Z(0-)=(0,0)$ (an arrival occurs to an empty system at time $t=0$), consider $\alpha(B_0)$, the age at the end of the first busy period, and let $\bar{\alpha}$ be the random variable having its distribution.

Central to our analysis is the following continuous-time Markov process that behaves like the interarrival age during idle periods of the workload process. Formally, we let $\{\Phi(t): t\geq 0\}$ be defined by the infinitesimal generator

$$G_\Phi f(\gamma)=\lim_{t\to 0}\frac{\mathbb{E}_\gamma f(\Phi(t))-f(\gamma)}{t}=f'(\gamma)+\eta(\gamma)\big(\mathbb{E}f(\bar{\alpha})-f(\gamma)\big),\quad \gamma\geq 0. \tag{EC.21}$$

This process grows at unit rate and, at its jump times, $\Phi(t)\overset{d}{=}\bar{\alpha}$. Jumps occur with rate $\eta(\Phi(t))$.

Let $\{\Phi(t): t\geq 0\}$ and $\{\widetilde{\Phi}(t): t\geq 0\}$ be two copies (possibly coupled) of the Markov process defined by (EC.21). The following technical lemma is proved in Appendix EC.5.2.3. The conditioning on the right-hand side of (EC.22) is non standard and means: pick a single $\bar{\alpha}$-distributed random variable $\xi$ and run $\{\Phi(t), t\geq 0\}$ from time zero starting at $\xi$, and run $\{\widetilde{\Phi}(t), t\geq 0\}$ from time $s$ starting at $\xi$.

Lemma EC.10. *Let $v=v(x)=x/\delta$. For any $(x,r_a)\in\mathbb{S}$ with $r_a<v$ and any $s\geq 0$,*

$$\begin{aligned}&\mathbb{E}_{x+\delta s,r_a}\eta(\alpha(B_0))-\mathbb{E}_{x,r_a}\eta(\alpha(B_0))\\ &=\mathbb{E}\big(\eta(\Phi(v-r_a+s))-\eta(\widetilde{\Phi}(v-r_a+s))|\Phi(0)=\widetilde{\Phi}(s),\Phi(0)\overset{d}{=}\bar{\alpha}\big),\end{aligned} \tag{EC.22}$$

*Furthermore if $\Phi(0)\overset{d}{=}\bar{\alpha}$, then the inter-jump times of $\{\Phi(t): t\geq 0\}$ are i.i.d. $\bar{I}$.*

The final ingredient for the proof of Lemma 16 is a coupling $\{(\widetilde{\Phi}(t),\Phi(t)): t\geq 0\}$, which allows us to bound the expression in (EC.22). Defining

$$\eta_m(x,y)=\min\{\eta(x),\eta(y)\},\quad \eta_\Delta(x,y)=\max\{\eta(x),\eta(y)\}-\min\{\eta(x),\eta(y)\},$$

we let $\{(\widetilde{\Phi}(t), \Phi(t)),\ t \geq 0\}$ have the same distribution as the Markov process defined by the generator

$$\begin{aligned} G_J f(\tilde{\gamma},\gamma) = \; & \partial_{\tilde{\gamma}} f(\tilde{\gamma},\gamma) + \partial_{\gamma} f(\tilde{\gamma},\gamma) \\ & + \eta_m(\tilde{\gamma},\gamma)\big(\mathbb{E} f(\bar{\alpha},\bar{\alpha}) - f(\tilde{\gamma},\gamma)\big) \\ & + \eta_\Delta(\tilde{\gamma},\gamma) 1(\eta(\gamma) < \eta(\tilde{\gamma}))\big(\mathbb{E} f(\bar{\alpha},\gamma) - f(\tilde{\gamma},\gamma)\big) \\ & + \eta_\Delta(\tilde{\gamma},\gamma) 1(\eta(\gamma) > \eta(\tilde{\gamma}))\big(\mathbb{E} f(\tilde{\gamma},\bar{\alpha}) - f(\tilde{\gamma},\gamma)\big). \end{aligned} \tag{EC.23}$$

Note that the marginal law of either component of this process is equivalent to the Markov process defined in (EC.21). Furthermore, when this process jumps it either couples (with rate $\eta_m(\tilde{\gamma},\gamma)$), or only one of the components jumps.

*Proof of Lemma 16* Fix $v = x/\delta \geq 0$ and $r_a < v$, and let $\tau_C = \inf\{t \geq s : \widetilde{\Phi}(t) = \Phi(t)\}$. It follows that

$$\begin{aligned} & |\mathbb{E}_{x+\delta s, r_a} \eta(\alpha(B_0)) - \mathbb{E}_{x,r_a} \eta(\alpha(B_0))| \\ = \; & \left|\mathbb{E}\big(\eta(\Phi(v - r_a + s)) - \eta(\widetilde{\Phi}(v - r_a + s)) | \Phi(0) = \widetilde{\Phi}(s), \Phi(0) \overset{d}{=} \bar{\alpha}\big)\right| \\ \leq \; & \overline{\eta}\mathbb{P}\big(\tau_C > v - r_a + s \big| \Phi(0) = \widetilde{\Phi}(s), \Phi(0) \overset{d}{=} \bar{\alpha}\big). \end{aligned} \tag{EC.24}$$

If $\widetilde{\Phi}(s) > \Phi(s)$, then $\eta_m(\Phi(t), \widetilde{\Phi}(t)) = \eta(\widetilde{\Phi}(t))$ for all $t \geq s$ until the first jump of $\{\widetilde{\Phi}(t) : t \geq 0\}$, at which point coupling occurs. Since $\widetilde{\Phi}(s) \overset{d}{=} \bar{\alpha}$, we know by Lemma EC.10 that the first jump after $s$ happens after $\bar{I}$ amount of time, implying that the probability of no jump on $(s, v - r_a + s]$ is at most

$$\mathbb{P}(\bar{I} > v - r_a). \tag{EC.25}$$

Alternatively, if $\widetilde{\Phi}(s) < \Phi(s)$, then $\tau_C$ corresponds to the first jump time after $s$ of $\{\Phi(t) : t \geq 0\}$. Unlike $\widetilde{\Phi}(s)$, it is not true that $\Phi(s)$ is distributed like $\bar{\alpha}$, because the latter process has been running since time zero with $\Phi(0) \overset{d}{=} \bar{\alpha}$. We will shortly prove that the probability that $\{\Phi(t) : t \geq 0\}$ does not jump on $(s, v - r_a + s]$ is at most

$$\mathbb{P}(\bar{I} > v - r_a)\mathbb{E}(N_\Phi(s) | \Phi(0) \overset{d}{=} \bar{\alpha}), \tag{EC.26}$$

where $N_\Phi(t)$ is the number of jumps made by $\{\Phi(t):t\geq 0\}$ on $[0,t]$. Adding (EC.25) and (EC.26) yields

$$\overline{\eta}\mathbb{P}\big(\tau_C > v-r_a+s\big|\Phi(0)=\widetilde{\Phi}(s)\stackrel{d}{=}\bar{\alpha}\big)\leq\overline{\eta}(1+\mathbb{E}(N_\Phi(s)|\Phi(0)\stackrel{d}{=}\bar{\alpha}))\mathbb{P}(\bar{I}>v-r_a).$$

Let $\{N_U(t):t\geq 0\}$ be the counting process of a zero-delayed renewal process whose inter-event times have the interarrival distribution $U$. It follows that

$$\mathbb{E}(N_\Phi(s)|\Phi(0)\stackrel{d}{=}\bar{\alpha})\leq\mathbb{E}N_U(s)\leq\lambda s+\lambda^2\mathbb{E}U^2,\quad s\geq 0,$$

where the first inequality is due to the hazard rate $\eta(x)$ being nonincreasing, and the second inequality follows from Lorden's inequality (Lorden 1970).

It remains to verify (EC.26). Let $T_0=0$ and $T_n$, $n\geq 1$, be the $n$th jump time of $\{\Phi(t):t\geq 0\}$. Also let $J_n=T_n-T_{n-1}$, $n\geq 1$, be the duration between the $n$th and $(n-1)$st jump. We know by Lemma EC.10 that $J_n\stackrel{d}{=}\bar{I}$ whenever $\Phi(0)\stackrel{d}{=}\bar{\alpha}$. It follows that (to simplify notation, all probabilities and expectations that follow are conditional on $\Phi(0)\stackrel{d}{=}\bar{\alpha}$)

$$\begin{aligned}\mathbb{P}(N_\Phi(v-r_a+s)-N_\Phi(s)=0)&=\mathbb{P}(J_{N_\Phi(s)+1}>v-r_a+s-T_{N_\Phi(s)})\\&=\sum_{n=0}^{\infty}\mathbb{P}(J_{n+1}>v-r_a+s-T_n,N_\Phi(s)=n).\end{aligned}$$

Since $\{N_\Phi(s)=n\}=\{T_n<s,J_{n+1}>s-T_n\}$, the right-hand side equals

$$\begin{aligned}\sum_{n=0}^{\infty}\mathbb{P}(J_{n+1}>v-r_a+s-T_n,T_n<s)&\leq\sum_{n=0}^{\infty}\mathbb{P}(J_{n+1}>v-r_a,T_n<s)\\&=\mathbb{P}(\bar{I}>v-r_a)\sum_{n=0}^{\infty}\mathbb{P}(T_n<s)\\&=\mathbb{P}(\bar{I}>v-r_a)\sum_{n=0}^{\infty}\mathbb{P}(N_\Phi(s)\geq n)\\&=\mathbb{P}(\bar{I}>v-r_a)\mathbb{E}N_\Phi(s),\end{aligned}$$

where the first equality is true because $J_{n+1}\stackrel{d}{=}\bar{I}$ and $J_{n+1}$ is independent of $T_n$. □

REMARK EC.2. Another family of hazard rates for which the coupling probability can be controlled consists of those uniformly bounded away from zero: $\eta(x)\geq\underline{\eta}>0$. In this case, $\eta_m(\tilde{\gamma},\gamma)\geq\underline{\eta}$, so coupling occurs after an exponentially distributed time with rate $\underline{\eta}$, leading to the bound

$$\mathbb{P}\big(\tau_C>v-r_a+s\big|\Phi(0)=\widetilde{\Phi}(s),\ \Phi(0)\stackrel{d}{=}\bar{\alpha}\big)\leq\overline{\eta}\,e^{-\underline{\eta}(v-r_a)}.$$

#### EC.5.2.2. Proof of Lemma 15.

LEMMA EC.11. *Suppose that $U$ has a bounded density. Then for any $h \in \mathcal{M}_2$ and any $x \geq 0$,*

$$\partial_x^2 \mathbb{E}\big(F_h(x+\delta S,U) - F_h(x,U)\big) = \mathbb{E}^S\big(\partial_x^2 \mathbb{E}^U F_h(x+\delta S,U) - \partial_x^2 \mathbb{E}^U F_h(x,U)\big).$$

*Proof of Lemma 15* Observe that

$$\begin{aligned}\delta \bar{F}_h'''(x) &= \lambda \partial_x^2 \mathbb{E}\big(F_h(x+\delta S,U) - F_h(x,U)\big)\\ &= \lambda \mathbb{E}^S\big(\partial_x^2 \mathbb{E}^U F_h(x+\delta S,U) - \partial_x^2 \mathbb{E}^U F_h(x,U)\big),\end{aligned}$$

where the first equality is due to (49) and the second is due to Lemma EC.11. Applying the expression for $\partial_x^2 \mathbb{E}^U F_h(\cdot,U)$ from Lemma 14 to the right-hand side yields the result. □

*Proof of Lemma EC.11* Though we do not assume $F_h(z)$ to be well defined, note that

$$\begin{aligned}&\partial_x \mathbb{E}\big(F_h(x+\delta S,U) - F_h(x,U)\big)\\ &= \lim_{\epsilon\to 0}\left(\frac{1}{\epsilon}\mathbb{E}\big(F_h(x+\epsilon+\delta S,U) - F_h(x+\delta S,U)\big) - \frac{1}{\epsilon}\mathbb{E}\big(F_h(x+\epsilon,U) - F_h(x,U)\big)\right)\\ &= \partial_x \mathbb{E} F_h(x+\delta S,U) - \partial_x \mathbb{E} F_h(x,U),\end{aligned}$$

so that by differentiating both sides with respect to $x$, we arrive at

$$\partial_x^2 \mathbb{E}\big(F_h(x+\delta S,U) - F_h(x,U)\big) = \partial_x^2 \mathbb{E} F_h(x+\delta S,U) - \partial_x^2 \mathbb{E} F_h(x,U).$$

By repeating the arguments used to prove Lemma 6, one can check that

$$\partial_x^2 \mathbb{E} F_h(x+\delta S,U) = \partial_x \mathbb{E}^S \partial_x \mathbb{E}^U F_h(x+\delta S,U).$$

Similarly,

$$\partial_x \mathbb{E}^S \partial_x \mathbb{E}^U F_h(x+\delta S,U) = \mathbb{E}^S \partial_x^2 \mathbb{E}^U F_h(x+\delta S,U)$$

follows from repeating the proof of Lemma 14. □

#### EC.5.2.3. Proof of Lemma EC.10.

Define $\ell_n$, $n \geq 0$, to be the end of the $n$th busy period. Namely, $\ell_0 = B_0$ and

$$\ell_n = \ell_{n-1} + I_{n-1} + B_n, \quad n \geq 1.$$

Letting $T_0 = 0$ and $T_{n+1} = T_n + I_n$, $n \geq 0$, so that the $n$th idle period occurs on the interval $[\ell_n, \ell_n + T_{n+1} - T_n)$, we define

$$\Gamma(t) = \alpha(\ell_n + t - T_n), \quad t \in [T_n, T_{n+1}),\ n \geq 0, \tag{EC.27}$$

to be the age of the interarrival process at the instant when the server has idled for exactly $t$ time units.

Recalling our synchronous coupling $\{Z^{(x)}(t) : t \geq 0\}$ introduced in (25) and the corresponding initial busy period $B_0^{(x)}$, it follows that

$$\alpha(B_0^{(x)}) = \Gamma(v), \quad x \geq 0,$$

because $B_0^{(x)}$ is precisely the instant when $\{Z(t) : t \geq 0\}$ idles for $v$ time units. Therefore, for any $(x, r_a) \in \mathbb{S}$ with $r_a < v$ and any $s \geq 0$,

$$\begin{aligned}\mathbb{E}_{x+\delta s, r_a}\eta(\alpha(B_0)) - \mathbb{E}_{x,r_a}\eta(\alpha(B_0)) &= \mathbb{E}_{0,r_a}\eta(\alpha(B_0^{(x+\delta s)})) - \mathbb{E}_{0,r_a}\eta(\alpha(B_0^{(x)})) \\ &= \mathbb{E}_{0,r_a}\big(\eta(\Gamma(v+s)) - \eta(\Gamma(v))\big),\end{aligned} \tag{EC.28}$$

where we recall that $\mathbb{E}_{x,r_a}(\cdot)$ is the expectation conditioned on $Z(0) = (x, r_a)$. To proceed, we now show that although $\{\Gamma(t) : t \geq 0\}$ is defined via (EC.27), it is also a Markov process with generator (EC.21).

We know from (EC.27) that $\{\Gamma(t) : t \geq 0\}$ increases at a unit rate. It jumps at times $T_{n+1}$, $n \geq 0$, and its distribution at jump times satisfies $\Gamma(T_{n+1}) = \alpha(\ell_{n+1}) \stackrel{d}{=} \bar{\alpha}$ for all $n \geq 0$, because $\ell_{n+1}$ marks the end of a busy period initiated by an arrival to an empty system. Lastly, conditioned on $\Gamma(t)$, the probability that a jump occurs on the interval $(t, t+dt)$ equals

$$\frac{\mathbb{P}(\Gamma(t) < U < \Gamma(t) + dt)}{\mathbb{P}(U > \Gamma(t))} = \eta(\Gamma(t))dt + o(dt),$$

where $o(dt) \to 0$ as $dt \to 0$; the Markov property is simple to verify. Thus $\{\Gamma(t) : t \geq 0\}$ is a Markov process described by the generator (EC.21) and, given the same initial condition, both $\{\Phi(t) : t \geq 0\}$ and $\{\Gamma(t) : t \geq 0\}$ are equal in distribution. We now argue that given $\Phi(0) \sim \bar{\alpha}$, the inter-jump times of $\{\Phi(t) : t \geq 0\}$ are i.i.d. $\bar{I}$.

By construction, the value of $\Gamma(\cdot)$ at jump times $T_{n+1}$ satisfies $\Gamma(T_{n+1}) \stackrel{d}{=} \bar{\alpha}$, and the inter-jump durations $T_{n+2} - T_{n+1} \stackrel{d}{=} \bar{I}$ for $n \geq 0$. Since $\{\Gamma(t) : t \geq 0\}$ is a Markov process, it follows that conditioned on $\Gamma(0) \stackrel{d}{=} \bar{\alpha}$, the time until the first jump has the same distribution as $\bar{I}$.

Coming back to right-hand side of (EC.28), suppose that $Z(0) = (0, r_a)$. Then $\Gamma(r_a) = \alpha(\ell_1) \stackrel{d}{=} \bar{\alpha}$, because $B_0 = 0$ and $I_0 = r_a$, which in turn yields $r_a \in [T_1, T_2)$ (since $T_1 = r_a$). Furthermore, conditioned on both $Z(0) = (0, r_a)$ and $\Gamma(r_a)$, the value of $\Gamma(r_a + t)$, $t \geq 0$, is independent of $Z(0)$. It follows that for any $v$ and $r_a < v$,

$$\begin{aligned}
\mathbb{E}_{0,r_a}\big(\eta(\Gamma(v+s)) - \eta(\Gamma(v))\big) &= \mathbb{E}\big(\eta(\Gamma(v+s)) - \eta(\Gamma(v)) | \Gamma(r_a) \stackrel{d}{=} \bar{\alpha}\big) \\
&= \mathbb{E}\big(\eta(\Phi(v+s)) - \eta(\Phi(v)) | \Phi(r_a) \stackrel{d}{=} \bar{\alpha}\big) \\
&= \mathbb{E}\big(\eta(\Phi(v - r_a + s)) - \eta(\Phi(v - r_a)) | \Phi(0) \stackrel{d}{=} \bar{\alpha}\big),
\end{aligned}$$

where the third equality comes from the Markov property. Finally, if $\{\widetilde{\Phi}(t) : t \geq 0\}$ is a copy of the Markov process defined by (EC.21), then

$$\begin{aligned}
&\mathbb{E}\big(\eta(\Phi(v - r_a + s)) - \eta(\Phi(v - r_a)) | \Phi(0) \stackrel{d}{=} \bar{\alpha}\big) \\
&= \mathbb{E}\big(\eta(\Phi(v - r_a + s)) - \eta(\widetilde{\Phi}(v - r_a)) | \Phi(0) = \widetilde{\Phi}(0), \Phi(0) \stackrel{d}{=} \bar{\alpha}\big) \\
&= \mathbb{E}\big(\eta(\Phi(v - r_a + s)) - \eta(\widetilde{\Phi}(v - r_a + s)) | \Phi(0) = \widetilde{\Phi}(s), \Phi(0) \stackrel{d}{=} \bar{\alpha}\big).
\end{aligned}$$

□

### EC.5.3. Proof of Lemma 12

*Proof of Lemma 12* We first prove (40). Let $\hat{h}(x) = x$, recall that $\bar{F}_{\hat{h}}(0) = 0$ due to (28), and consider the Poisson equation (33) with $h(x) = \hat{h}(x)$ evaluated at $x = 0$, which results in

$$\lambda \mathbb{E}\big(F_{\hat{h}}(\delta S, U) - F_{\hat{h}}(0, U)\big) = \delta \mathbb{E} V,$$

Using the synchronous coupling $\{Z^{(\epsilon)}(t)\}$ introduced in Section 3.3, it follows that

$$\delta \mathbb{E}V = \lambda \mathbb{E}\Big( \int_0^{B^{(\delta S)}} \mathbb{E}_{0,U}\big(X^{(\delta S)}(t) - X(t)\big) dt \Big). \tag{EC.29}$$

From (25) we know that the difference $X^{(\delta S)}(t) - X(t)$ decays at rate $\delta$ only during the idle periods of $\{X(t) : t \geq 0\}$. Recall that the idle and busy period durations are $I_0, I_1, \ldots$, and $B_0, B_1, B_2, \ldots$, respectively, and that $B_0 = 0$ because $X(0) = 0$. It follows that for all times $t$ corresponding to the busy period $B_k$, $k \geq 1$,

$$X^{(\delta S)}(t) - X(t) = \delta \Big( S - \sum_{i=0}^{k-1} I_i \Big)^+.$$

Letting $\mathcal{I} = \{t \in [0, B^{(\delta S)}] : X(t) = 0\}$, it follows that

$$\begin{aligned} \int_0^{B^{(\delta S)}} \big(X^{(\delta S)}(t) - X(t)\big) dt &= \int_{\mathcal{I}} \big(X^{(\delta S)}(t) - X(t)\big) dt + \int_{[0,B^{(\delta S)}]\setminus \mathcal{I}} \big(X^{(\delta S)}(t) - X(t)\big) dt \\ &= \delta S^2/2 + \sum_{k=1}^{\infty} B_k \delta \Big( S - \sum_{i=0}^{k-1} I_i \Big)^+. \end{aligned}$$

We conclude (40) by combining this equation with (EC.29), noting that $B_k$ is independent of $(S - I_0 - I_1 \ldots - I_{k-1})^+$ and $B_k \stackrel{d}{=} \bar{B}$, and that $I_0 \stackrel{d}{=} U$ since $Z(0) = (0, U)$. The equality in (41) is true because $\mathbb{E}V = \lambda \mathbb{E}S^2/2 + \rho \mathbb{E}W$ due to Corollary X.3.5 of Asmussen (2003). The inequality follows from the well-known bound in (7') of Kingman (1962), which says that

$$\mathbb{E}W \leq \frac{\mathrm{Var}(S-U)}{2\mathbb{E}(S-U)} = \frac{\rho \mathrm{Var}(S-U)}{2(1-\rho)\mathbb{E}S}.$$

□

### EC.5.4. Proof of Lemma 13

For convenience, let $\nu = 1/\mathbb{E}Y_2$. We use the following inequalities throughout the proof, which follow from the fact that $Y_2$ has density $\nu e^{-\nu y}$ and is independent of $R_a$.

$$\mathbb{P}(\delta R_a \geq Y_2) = \mathbb{E}(1 - e^{-\nu \delta R_a}) \leq \nu \delta \mathbb{E}R_a, \tag{EC.30}$$

$$\mathbb{E}Y_2 1(\delta R_a \geq Y_2) = \frac{1}{\nu} \mathbb{E}\big((1 - e^{-\nu\delta R_a}) - \nu \delta R_a e^{-\nu \delta R_a}\big) \leq 2\nu\delta^2 \mathbb{E}R_a^2, \tag{EC.31}$$

$$\mathbb{E}Y_2^2 1(\delta R_a \geq Y_2) = \frac{1}{\nu^2} \mathbb{E}\big(2(1 - e^{-\nu\delta R_a}) - 2\nu\delta R_a e^{-\nu\delta R_a} - (\nu \delta R_a)^2 e^{-\nu\delta R_a}\big) \leq 3\nu\delta^3 \mathbb{E}R_a^3. \tag{EC.32}$$

We also use the facts that $U, S, S', R_a, \bar{I}$, and $Y_2$ are independent, and that $S$ and $S'$ have the same distribution. We first prove (42)–(44) and then (45)–(46).

**Proof of (42)–(44).** Inequality (42) follows from $J(x,r_a) = -(x\wedge \delta r_a) + \delta S'$. Next, we prove (43). We first note that the definition of $\epsilon(x,r_a,s)$ in Lemma 8 and the fact that $h(x) = x$ both imply that

$$\int_0^{r_a} \big| h\big((x+\delta s-\delta t)^+\big) - h\big((x-\delta t)^+\big)\big|\, dt \le r_a \delta s.$$

Furthermore,

$$\begin{aligned}
&1(\delta r_a > x)\mathbb{E}^{S'}\Big(\int_{-x}^{-(x+\delta s)} \partial_x \mathbb{E}^U F_h(x+\delta s+v+\delta S', U)dv\Big)\\
&\le 1(\delta r_a > x)\delta s \mathbb{E}^{S'}\Big(\sup_{0\le w\le \delta s} \partial_x \mathbb{E}^U F_h(w+\delta S', U)\Big)\\
&\le 1(\delta r_a > x) s(\delta s + \delta \mathbb{E}S')(1+2\overline{\eta}\mathbb{E}\bar{B}),
\end{aligned}$$

where in the second inequality we used (36) of Lemma 10. Combining the bounds and using (EC.30) yields

$$\begin{aligned}
\mathbb{E}\left|\epsilon(Y_2,R_a,S)\right| &\le \delta \mathbb{E}R_a\mathbb{E}S + \mathbb{P}(\delta R_a \ge Y_2)\delta(\mathbb{E}S^2 + (\mathbb{E}S)^2)(1+2\overline{\eta}\mathbb{E}\bar{B})\\
&\le \delta \mathbb{E}R_a\mathbb{E}S + \delta^2 \nu \mathbb{E}R_a(\mathbb{E}S^2 + (\mathbb{E}S)^2)(1+2\overline{\eta}\mathbb{E}\bar{B}).
\end{aligned}$$

Next we prove (44). Recall from Lemma 10 that

$$\begin{aligned}
&\big|\delta \bar{F}_h'(x)\big| \le x(1+(\lambda+\overline{\eta})\mathbb{E}\bar{B}),\\
&\big|\delta \bar{F}_h''(x)\big| \le (1+x)\big(1+(\lambda+\overline{\eta})\mathbb{E}\bar{B}\big), \quad x\ge 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\big|1(\delta R_a\ge Y_2)\big(\delta\bar{F}_h'(\delta S) - \delta \bar{F}_h'(Y_2) - \delta(S-R_a)\delta\bar{F}_h''(Y_2)\big)\big|\\
&\le \Big(\mathbb{E}\big(1(\delta R_a\ge Y_2)(\delta S+Y_2)\big) + \mathbb{E}\big(1(\delta R_a \ge Y_2)\delta(S-R_a)(1+Y_2)\big)\Big)\big(1+(\lambda+\overline{\eta})\mathbb{E}\bar{B}\big)\\
&\le \delta^2\big(\nu\big(2\mathbb{E}S\mathbb{E}R_a + 5\mathbb{E}R_a^2\big) + \mathbb{E}R_a^2\big)\big(1+(\lambda+\overline{\eta})\mathbb{E}\bar{B}\big).
\end{aligned}$$

The last inequality follows from using (EC.30) and (EC.31) to show that

$$\begin{aligned}
&\mathbb{E}\big(1(\delta R_a \geq Y_2)(\delta S + Y_2)\big) \leq \delta^2 \nu \mathbb{E}S\mathbb{E}R_a + 2\nu\delta^2\mathbb{E}R_a^2,\\
&\delta\mathbb{E}S\mathbb{E}\big(1(\delta R_a \geq Y_2)(1+Y_2)\big) \leq \delta\mathbb{E}S\big(\nu\delta\mathbb{E}R_a + 2\nu\delta^2\mathbb{E}R_a^2\big),\\
&\delta\mathbb{E}\big(1(\delta R_a \geq Y_2)R_a(1+Y_2)\big) \leq \delta\big(\nu\delta\mathbb{E}R_a^2 + \delta\mathbb{E}R_a^2\big),
\end{aligned}$$

where in the final inequality we used the fact that $\mathbb{E}\big(1(\delta R_a \geq Y_2)R_a Y_2\big) \leq \delta\mathbb{E}R_a^2$ instead of (EC.31). Using the latter would have resulted in a term involving $\mathbb{E}R_a^3$.

**Proof of (45)–(46).** Recall that

$$C = (3 + (\mathbb{E}U + \mathbb{E}\bar{I})\lambda(1+\rho+\lambda^2\mathbb{E}U^2))\nu\overline{\eta}\delta\mathbb{E}\bar{B}.$$

We claim that

$$\mathbb{E}\big|\delta^2\bar{F}_h'''(Y_2+u)\big| \leq \overline{C}, \qquad u>0, \tag{EC.33}$$

$$\mathbb{E}^{Y_2}\Big(1(\delta R_a \leq Y_2)\big|\delta^2\bar{F}_h'''(Y_2+u)\big|\Big) \leq \overline{C}, \qquad u \in [-\delta R_a, 0]. \tag{EC.34}$$

Then (45) follows by applying the bound in (EC.33) to

$$\lambda\mathbb{E}\Big(\int_0^{\delta S}(\delta S - v)\int_0^v \big|\bar{F}_h'''(Y_2+u)\big|\,dudv\Big) = \lambda\mathbb{E}^S\Big(\int_0^{\delta S}(\delta S - v)\int_0^v \mathbb{E}^{Y_2}\big|\bar{F}_h'''(Y_2+u)\big|\,dudv\Big).$$

To prove (46), observe that both (EC.33) and (EC.34) imply that

$$\begin{aligned}
&\delta\frac{1}{\delta^2}\mathbb{E}\Big(\int_0^{\delta(S-R_a)}\int_0^v \big|1(\delta R_a < Y_2)\delta^2\bar{F}_h'''(Y_2+u)\big|dudv\Big)\\
&= \delta\frac{1}{\delta^2}\mathbb{E}^{S,R_a}\Big(\int_0^{\delta(S-R_a)}\int_0^v \mathbb{E}^{Y_2}\big|1(\delta R_a < Y_2)\delta^2\bar{F}_h'''(Y_2+u)\big|dudv\Big)\\
&\leq \delta\mathbb{E}(S-R_a)^2\overline{C}.
\end{aligned}$$

It remains to prove (EC.33) and (EC.34). Recall from Lemma 11 that for any $x,u$ such that $x+u \geq 0$,

$$\big|\delta^2\bar{F}_h'''(x+u)\big| \leq \big(\mathbb{P}(\delta U > x+u)3\lambda + \mathbb{P}(\delta U < x+u < \delta\bar{I} + \delta U)\lambda(1+\rho+\lambda^2\mathbb{E}U^2)\big)\overline{\eta}\mathbb{E}\bar{B}. \tag{EC.35}$$

Then (EC.33) follows once we observe that for any $u>0$,

$$\mathbb{P}(\delta U > Y_2 + u) \leq \mathbb{P}(\delta U > Y_2) = \mathbb{E}(1 - e^{-\nu\delta U}) \leq \nu\delta\mathbb{E}U,$$
$$\mathbb{P}(\delta U < Y_2 + u < \delta\bar{I} + \delta U) \leq \nu\delta(\mathbb{E}U + \mathbb{E}\bar{I}), \tag{EC.36}$$

where the last inequality is true because

$$\begin{aligned}
&\mathbb{P}(\delta U < Y_2 + u < \delta\bar{I} + \delta U) \\
&= \mathbb{P}(Y_2 \leq \delta U < Y_2 + u < \delta\bar{I} + \delta U) + \mathbb{P}(\delta U < Y_2, Y_2 + u < \delta\bar{I} + \delta U) \\
&\leq \mathbb{P}(\delta U \geq Y_2) + \mathbb{P}(\delta U < Y_2 < \delta U + \delta\bar{I}) \\
&= \mathbb{E}(1 - e^{-\nu\delta U}) + \mathbb{E}e^{-\nu\delta U}(1 - e^{-\nu\delta\bar{I}}) \leq \nu\delta(\mathbb{E}U + \mathbb{E}\bar{I}).
\end{aligned}$$

Similarly, (EC.34) follows from the fact that for any $u \in [-\delta R_a, 0]$,

$$\mathbb{E}^{Y_2,U}\big(1(\delta R_a < Y_2)1(\delta U > Y_2 + u)\big) \leq \nu\delta\mathbb{E}U,$$
$$\mathbb{E}^{Y_2,U,\bar{I}}\big(1(\delta U < Y_2 + u < \delta\bar{I} + \delta U)\big) \leq \nu\delta(\mathbb{E}U + \mathbb{E}\bar{I})\mathbb{E}\bar{B}. \tag{EC.37}$$

The first inequality in (EC.37) follows from

$$\begin{aligned}
&\mathbb{E}^{Y_2,U}\big(1(\delta R_a < Y_2)1(\delta U > Y_2 + u)\big) \\
&\leq \mathbb{E}^{Y_2,U}\big(1(\delta R_a < Y_2 < \delta U + \delta R_a)\big) = e^{-\nu\delta R_a}\mathbb{E}(1 - e^{-\nu\delta U}) \leq \nu\delta\mathbb{E}U
\end{aligned}$$

and the second from

$$\begin{aligned}
E^{Y_2,U,\bar{I}}\big(1(\delta U < Y_2 + u < \delta\bar{I} + \delta U)\big) &\leq E^{Y_2,U,\bar{I}}\big(1(\delta U - u < Y_2 < \delta\bar{I} + \delta U - u)\big) \\
&= \mathbb{E}\big(e^{-\nu(\delta U - u)}(1 - e^{-\nu\delta\bar{I}})\big) \leq \nu\delta\mathbb{E}\bar{I}.
\end{aligned}$$

□

## EC.6. The $G/M/\infty$ system: supporting proofs

To prove Proposition EC.1 we require the following inversion formula, which is proved exactly like Lemma 3.

Lemma EC.12. *Initialize $Z(0) \sim Z$ and fix $f : \mathbb{R} \to \mathbb{R}$ with*

$$\mathbb{E}\,|f(X)| < \infty \quad \textit{and} \quad \mathbb{E}\left|\int_0^{R_a(0)} f(X(t))dt\right| < \infty,$$

*which holds, in particular, when $f(x)$ is bounded. Then*

$$\mathbb{E}f(X) = \mathbb{E}\int_0^1 \int_0^{U(t)} f(X(t+u))dudA(t) \tag{EC.38}$$

We now have what we need to prove Proposition EC.1.

*Proof of Proposition EC.1* Repeating the arguments used in the proof of Proposition 1 yields

$$\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dA(t) = \frac{1}{2}\delta^2\lambda c_U^2\mathbb{E}f''(X) + \epsilon_A(f) = \frac{1}{2}\mu c_U^2\mathbb{E}f''(X) + \epsilon_A(f),$$

where

$$\begin{aligned}\epsilon_A(f) = {} & \frac{1}{6}\delta^3\mathbb{E}\int_0^1 (1-\lambda U(t))^3 f'''(\xi(t))dA(t) \\ & - \frac{1}{2}\delta^2\lambda c_U^2\mathbb{E}\int_0^1 \int_0^{U(t)} \big(X(t+u) - X(t-)\big) f'''(\xi(t+u))dudA(t).\end{aligned} \tag{EC.39}$$

To bound the first term on the right-hand side, we use the fact that $\mathbb{E}A(1) = \lambda$ and $\delta^2\lambda = \mu$ to get

$$\frac{1}{6}\delta^3\mathbb{E}\int_0^1 \big|(1-\lambda U(t))^3 f'''(\xi(t))\big|\, dA(t) \le \frac{1}{6}\delta\mu\|f'''\|\mathbb{E}\,|1-\lambda U|^3$$

To bound the second term in (EC.39), let $D[t, t+u]$ be the number of departures on $[t, t+u]$ and recall that $X(t) = \delta(Q(t) - R)$. It follows that

$$\begin{aligned}& \mathbb{E}\int_0^1 \int_0^{U(t)} \big|\big(X(t+u) - X(t-)\big) f'''(\xi(t+u))\big|\, dudA(t) \\ \le {} & \|f'''\|\delta\mathbb{E}\int_0^1 \int_0^{U(t)} |Q(t+u) - Q(t-)|\, dudA(t) \\ \le {} & \|f'''\|\delta\mathbb{E}\int_0^1 \int_0^{U(t)} (1 + D[t, t+u])dudA(t) \\ \le {} & \|f'''\|\delta\mathbb{E}\int_0^1 \int_0^{U(t)} (1 + \mu Q(t)u)dudA(t)\end{aligned}$$

$$
\begin{aligned}
&= \|f'''\|\delta\mathbb{E}\int_0^1 (U(t)+\mu Q(t)U^2(t)/2)dA(t) \\
&= \|f'''\|\delta\Big(1+\frac{1}{2}\mathbb{E}U^2\mu\mathbb{E}\int_0^1 Q(t)dA(t)\Big) \\
&= \|f'''\|\delta\Big(1+\frac{1}{2}\mathbb{E}U^2\mu\mathbb{E}\int_0^1 (1+Q(t-))dA(t)\Big).
\end{aligned}
$$

The third inequality is true because the number of departures on $[t, t+u]$ is bounded by a Poisson process with rate $\mu Q(t)$. Using $\mathbb{E}A(1)=\lambda$ and (EC.8) of Lemma EC.2, the right-hand side equals

$$
\begin{aligned}
&\|f'''\|\delta\Big(1+\frac{1}{2}\mathbb{E}U^2\mu(\lambda+\mu\mathbb{E}(Q-R)^2+\lambda R-\lambda)\Big) \\
&= \|f'''\|\delta\Big(1+\frac{1}{2}\lambda^2\mathbb{E}U^2(\mathbb{E}(Q-R)^2/R^2+1)\Big).
\end{aligned}
$$

We arrive at the stated bound on $|\epsilon_A(f)|$.

We now prove (EC.4). Taylor expansion yields

$$
\mu Q\big(f(\widetilde{X}-\delta)-f(\widetilde{X})\big) = -\mu\delta Q f'(\widetilde{X})+\frac{1}{2}\mu\delta^2 Q f''(\widetilde{X})+\frac{1}{6}\mu\delta^3 Q f'''(\xi).
$$

Since $\mu R=\lambda$, $X=\delta(Q-R)$, and $\widetilde{X}-X=-\delta\lambda R_a$, the first two terms on the right-hand side satisfy

$$
\begin{aligned}
-\mu\delta Q f'(\widetilde{X}) &= -\mu\delta R f'(\widetilde{X})-\mu\delta(Q-R)f'(\widetilde{X}) \\
&= -\mu\delta R f'(\widetilde{X})-\mu\delta(Q-R)f'(X)-\mu\delta(Q-R)(f'(\widetilde{X})-f'(X)) \\
&= -\delta\lambda f'(\widetilde{X})-\mu X f'(X)+\mu X\delta\lambda R_a f''(\xi),
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{2}\mu\delta^2 Q f''(\widetilde{X}) &= \frac{1}{2}\mu\delta^2 R f''(X)+\frac{1}{2}\mu\delta^2(Q-R)f''(X)+\frac{1}{2}\mu\delta^2 Q(f''(\widetilde{X})-f''(X)) \\
&= \frac{1}{2}\mu f''(\widetilde{X})+\frac{1}{2}\mu\delta X f''(X)-\frac{1}{2}\mu\delta^3 Q\lambda R_a f'''(\xi),
\end{aligned}
$$

implying that (EC.40) holds with

$$
\epsilon_D(f)=\mu\delta\mathbb{E}\big(X\lambda R_a f''(\xi)\big)-\frac{1}{2}\mu\delta^3\mathbb{E}\big(Q\lambda R_a f'''(\xi)\big)+\frac{1}{6}\mu\delta^3\mathbb{E}\big(Q f'''(\xi)\big). \tag{EC.40}
$$

To conclude, we recall that $\mathbb{E}Q = R$ and bound each of the three terms on the right-hand side as follows:

$$\begin{aligned}\mu\delta\mathbb{E}\left|X\lambda R_a f''(\xi)\right| &\leq \mu\delta\|f''\|\sqrt{\mathbb{E}X^2}\sqrt{\lambda^2\mathbb{E}R_a^2} = \mu\delta\|f''\|\sqrt{\mathbb{E}X^2}\sqrt{\lambda^3\mathbb{E}U^3/3},\\ \frac{1}{2}\mu\delta^3\mathbb{E}\big(Q\lambda R_a f'''(\xi)\big) &\leq \frac{1}{2}\mu\|f'''\|\delta^3\mathbb{E}\big(Q\lambda R_a\big)\\ \frac{1}{6}\mu\delta^3\mathbb{E}\big(Qf'''(\xi)\big) &\leq \frac{1}{6}\mu\|f'''\|\delta^3\mathbb{E}Q = \frac{1}{6}\mu\|f'''\|\delta.\end{aligned}$$

□

#### EC.6.1. Proof of Lemma EC.2

The first two equalities in (EC.6) are argued as in Lemma 2 and the third is obtained from the BAR (EC.1) with $f(z) = q\wedge M$. For the relationship in (EC.7), we invoke the BAR with $f(z) = ((q\wedge M) - \lambda r_a)^2$ to get

$$\begin{aligned}0 =&\ 2\lambda\mathbb{E}\big(Q\wedge M - \lambda R_a\big) + \mu\mathbb{E}\Big(Q1(Q\leq M)\big((Q-1-\lambda R_a)^2 - (Q-\lambda R_a)^2\big)\Big)\\ &+\mathbb{E}\int_0^1 1(Q(t-)\leq M-1)\big((Q(t-)+1-\lambda U(t))^2 - (Q(t-))^2\big)dA(t)\\ =&\ 2\lambda\mathbb{E}\big(Q\wedge M - \lambda R_a\big) + \mu\mathbb{E}\big(Q1(Q\leq M)\big(-2(Q-\lambda R_a)+1\big)\big)\\ &+\mathbb{E}\int_0^1 1(Q(t-)\leq M-1)\big(2Q(t-)(1-\lambda U(t)) + (1-\lambda U(t))^2\big)dA(t)\end{aligned}$$

Taking $M\to\infty$ yields

$$0 = 2\lambda\mathbb{E}Q - 2\lambda^2\mathbb{E}R_a - 2\mu\mathbb{E}Q^2 + 2\mu\lambda\mathbb{E}QR_a + \mu\mathbb{E}Q + \mathbb{E}(1-\lambda U)^2\mathbb{E}A(1).$$

Divide by $2\mu$, move $\mathbb{E}Q^2$ to the left-hand side, and use $\mathbb{E}Q = R$ to get

$$\begin{aligned}\mathbb{E}Q^2 &= \mathbb{E}(Q-R)^2 + R^2\\ &= R^2 - R\lambda^2\mathbb{E}U^2/2 + \lambda\mathbb{E}QR_a + R/2 + R\mathbb{E}(1-\lambda U)^2/2,\end{aligned}$$

implying that

$$\begin{aligned}\mathbb{E}(Q-R)^2 &= -R\lambda^2\mathbb{E}U^2/2 + \lambda\mathbb{E}QR_a + R/2 + R(1-2+\lambda^2\mathbb{E}U^2)/2\\ &= \lambda\mathbb{E}QR_a\\ &= \lambda\mathbb{E}(Q-R)R_a + R\lambda^2\mathbb{E}U^2/2\\ &\leq \lambda\sqrt{\mathbb{E}(Q-R)^2}\sqrt{\mathbb{E}R_a^2} + R\lambda^2\mathbb{E}U^2/2.\end{aligned}$$

Since $x^2 \le bx + c$ implies that $x^2 \le (b^2 + x^2)/2 + c$ and, therefore, $x^2 \le b^2 + 2c$, and since $\mathbb{E}R_a^2 = \lambda \mathbb{E}U^3/3$, we obtain (EC.7).

To prove (EC.8) we use $f(z) = (q \wedge m)^2$ in the BAR to get

$$\begin{aligned}
0 = \mu\mathbb{E}\big(Q1(Q \le M)(-2Q+1)\big) + \mathbb{E}\int_0^1 1(Q(t-) \le M-1)(2Q(t-)+1)dA(t) \\
= -2\mu\mathbb{E}(Q^2 1(Q \le M)) + \mu E(Q1(Q \le M)) \\
+ 2\mathbb{E}\int_0^1 1(Q(t-) \le M-1)Q(t-)dA(t) + \mathbb{E}\int_0^1 1(Q(t-) \le M-1)dA(t).
\end{aligned}$$

Taking $M \to \infty$ and using $\mathbb{E}Q = R$ we arrive at

$$\begin{aligned}
2\mathbb{E}Q^2 &= \mathbb{E}Q + \frac{2}{\mu}\mathbb{E}\int_0^1 Q(t-)dA(t) + \frac{1}{\mu}\mathbb{E}A(1) \\
&= R + \frac{2}{\mu}\mathbb{E}\int_0^1 Q(t-)dA(t) + R.
\end{aligned}$$

Using $\mathbb{E}A(1) = \lambda$ and rearranging terms, we get

$$\frac{1}{\mu}\mathbb{E}\int_0^1 Q(t-)dA(t) = \mathbb{E}Q^2 - R = \mathbb{E}(Q-R)^2 + R^2 - R.$$

$\square$

### EC.6.2. Stein factor bounds for the Normal distribution

*Proof of Lemma EC.3* Given $h \in \mathrm{Lip}(1)$, define $\bar{h}(x) = h(\sigma x)/\sigma$ and $\bar{f}(x) = f_{\bar{h},1}(x)$. Let us verify that $f_{h,\sigma}(x) = \sigma \bar{f}(x/\sigma)$ solves the Poisson equation. Since

$$f'_{h,\sigma}(x) = \bar{f}'(x/\sigma) \quad \text{and} \quad f''_{h,\sigma}(x) = \bar{f}''(x/\sigma)/\sigma,$$

it follows that

$$\begin{aligned}
-xf'_{h,\sigma}(x) + \frac{1}{2}\sigma^2 f''_{h,\sigma}(x) &= \sigma\big(-\frac{x}{\sigma}\bar{f}'(x/\sigma) + \frac{1}{2}\bar{f}''(x/\sigma)\big) \\
&= \frac{\sigma}{\mu}(\mathbb{E}\bar{h}(Y/\sigma) - \bar{h}(x/\sigma)) = \frac{1}{\mu}(\mathbb{E}h(Y) - h(x)),
\end{aligned}$$

and our Stein factor bounds follow from the standard result (Chen et al. 2011, Lemma 2.4) that:

$$\|\bar{f}'\| \le \frac{2}{\mu}, \quad \|\bar{f}''\| \le \frac{\sqrt{2/\pi}}{\mu}, \quad \text{and} \quad \|\bar{f}'''\| \le \frac{2}{\mu}.$$

$\square$

## EC.7. The JSQ system: supporting proofs

Let $A(t)$ and $U(t)$ be defined as in Section 2.1. Similarly, let $D_i(t)$ denote the number of departures from server $i$ on $[0,t]$ and if a departure occurs from server $i$ at time $t$, we let $S_i(t)$ be the service time of the next customer to be served by that server. We also let $\Lambda_i(t)$ denote the (random) remaining time until a customer gets routed to server $i$. The following BAR is proved just like Lemma 1.

Lemma EC.13. *Initialize $Z(0)\sim Z$. If $f(Z(s))$ satisfies the FTC conditions with probability one under $Z(0)\sim Z$ and if the integrability condition*

$$\mathbb{E}\left|f(Z)\right|,\ \mathbb{E}\left|\partial_{r_a} f(Z)\right|,\ \mathbb{E}\left|\partial_{r_{s,i}} f(Z)\right|,\ \mathbb{E}\int_0^t \left|\Delta f(Z(s-))\right| dD_i(s),\ \mathbb{E}\int_0^t \left|\Delta f(Z(s-))\right| dA(s) < \infty$$

*holds, then*

$$\begin{aligned}
&-\mathbb{E}(\partial_{r_a} f(Z)) - \sum_{i=1}^n \mathbb{E}(1(Q_i>0)\partial_{r_{s,i}} f(Z))\\
&+\mathbb{E}\int_0^1 \Delta f(Z(t-))dA(t) + \sum_{i=1}^n \mathbb{E}\int_0^1 \Delta f(Z(t-))dD_i(t) = 0.
\end{aligned} \tag{EC.41}$$

Define the compensated total customer count

$$\widetilde{X}(t) = X(t) - \delta\lambda R_a(t) + \sum_{i=1}^n \delta\mu R_{s,i}(t) \quad \text{and} \quad \widetilde{X} = X - \delta\lambda R_a + \sum_{i=1}^n \delta\mu R_{s,i}.$$

Specialized to $\widetilde{X}$, the BAR (EC.41) becomes

$$\begin{aligned}
&\delta\mathbb{E}\Big(\Big(\lambda - \sum_{i=1}^n 1(Q_i>0)\mu\Big) f'(\widetilde{X})\Big) + \mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dA(t)\\
&+\sum_{i=1}^n \mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dD_i(t) = 0.
\end{aligned} \tag{EC.42}$$

The following is an analog of Proposition 1, and is proved in Appendix EC.7.1.

Proposition EC.2. *If $f\in C^2(\mathbb{R})$ with $f''(x)$ absolutely continuous and $\|f''\|, \|f'''\|<\infty$, then, provided that all expectations are well defined,*

$$\delta\mathbb{E}\Big(\Big(\lambda - \sum_{i=1}^n 1(Q_i>0)\mu\Big) f'(\widetilde{X})\Big) = -n\mu\delta^2\mathbb{E}f'(X) + n\mu\delta^2 f'(0) + \epsilon_0(f), \tag{EC.43}$$

$$\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dA(t) = \frac{1}{2}\delta^2\lambda c_U^2\mathbb{E}f''(X) + \epsilon_A(f), \tag{EC.44}$$

$$\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-))dD_i(t) = \frac{1}{2}\delta^2\mu c_S^2\mathbb{E}f''(X) + \epsilon_{D,i}(f), \tag{EC.45}$$

*where*

$$\begin{aligned}
|\epsilon_0(f)| \le{}& \frac{1}{2}n\mu\delta^3\|f''\|\big(\lambda^2\mathbb{E}U^2/2 + (\lambda\mu\mathbb{E}S^2/2 + \mu\delta\mathbb{E}S)\big)\\
&+ \delta^2\mu\|f''\|\sum_{i=1}^n \mathbb{E}\Big(1(Q_i=0)\Big(\sum_{j=1}^n Q_j + \lambda R_a + \sum_{j=1}^n \mu R_{s,j}\Big)\Big),\\
|\epsilon_A(f)| \le{}& \frac{1}{2}\delta^3\|f'''\|\Big(\frac{1}{3}\lambda\mathbb{E}\,|1-\lambda U|^3 + c_U^2 n\mu\lambda(\rho(\lambda/n)\mathbb{E}S^2/2 + \delta\mathbb{E}S + \lambda\mathbb{E}U^2/2)\\
&\qquad + \lambda c_U^2(1+n+\lambda\mu\mathbb{E}U^2 + \mu^2\mathbb{E}S^2)\Big),\\
|\epsilon_{D,i}(f)| \le{}& \frac{1}{2}\delta^3\|f'''\|\bigg(\frac{1}{3}\frac{\lambda}{n}\mathbb{E}\,|1-\mu S|^3\\
&+ c_S^2\big((\lambda^2\mu\rho + 2(n-1)\mu^2\rho\lambda)\mathbb{E}S^2/2n + \delta(\lambda\mu + 2(n-1)\mu^2)\mathbb{E}S + \frac{1}{2}\lambda^2\mu\mathbb{E}U^2\big)\\
&+ \mu c_S^2 n\rho + \mu c_S^2(\lambda/n)\big(\mathbb{E}S + \lambda\mathbb{E}S^2 + \lambda^2\mathbb{E}U^2\mathbb{E}S\big)\\
&+ c_S^2\lambda\big(1 + (\lambda\mu\mathbb{E}U^2/2 + 1/\rho) + 2\mu^2\mathbb{E}S^2\big)\bigg) + \frac{1}{2}\delta^3\|f''\|c_S^2\mu\lambda/n.
\end{aligned}$$

The expansions in (EC.43)–(EC.45) suggest the diffusion generator

$$G_Y f(x) = -n\mu\delta^2 f'(x) + \frac{1}{2}\delta^2(\lambda c_U^2 + n\mu c_S^2)f''(x) + n\mu\delta^2 f'(0), \quad x \ge 0,$$

which, like in Section 2, corresponds to the exponential distribution with mean $(\rho c_U^2 + c_S^2)/2$. We now combine everything to prove Theorem 3.

*Proof of Theorem 3* Fix $h \in \mathrm{Lip}(1)$. Since $\mathbb{E}X^2 < \infty$, we can verify that $f_h(\widetilde{X})$ satisfies the conditions of the BAR (EC.41) in the same way as we did in the proof of Theorem 1. For the Stein factor bounds, we use Lemma 4 with $\theta = n\mu\delta^2$ and $\sigma^2 = \delta^2(\lambda c_U^2 + n\mu c_S^2)/2$ there to get

$$\|f_h''\| \le \frac{1}{n\mu\delta^2} \quad \text{and} \quad \|f_h'''\| \le \frac{4}{\delta^2(\lambda c_U^2 + n\mu c_S^2)}.$$

Combining these with the bounds on $|\epsilon_0(f_h)|$, $|\epsilon_A(f_h)|$, and $|\epsilon_{D,i}(f_h)|$ in Proposition EC.2 concludes the proof. □

### EC.7.1. Expanding the BAR

To prove Proposition EC.2 we require two auxiliary lemmas. The first lemma contains some useful relationships derived using the BAR.

LEMMA EC.14. *For any $1 \le i \ne j \le n$ and $m > 1$,*

$$\mathbb{E}A(1) = n\mathbb{E}D_i(1) = \lambda, \quad \mathbb{P}(Q_i > 0) = \rho, \tag{EC.46}$$

$$\mathbb{E}R_a^{m-1} = \lambda \mathbb{E}U^m/m, \quad \mathbb{E}(R_{s,i}^{m-1}1(Q_i > 0)) = (\lambda/n)\mathbb{E}S^m/m, \quad \mathbb{E}(R_{s,i}^m|Q_i = 0) = \mathbb{E}S^m, \tag{EC.47}$$

$$\mathbb{E}\int_0^1 R_a(t)dD_i(t) \le \mu(\mathbb{E}R_{s,i} + \mathbb{E}R_a), \quad \mathbb{E}\int_0^1 R_{s,i}(t)dA(t) \le \lambda(\mathbb{E}R_{s,i} + \mathbb{E}R_a), \tag{EC.48}$$

$$\mathbb{E}\int_0^1 R_{s,i}(t)dD_j(t) \le \mu(\mathbb{E}R_{s,i} + \mathbb{E}R_{s,j}). \tag{EC.49}$$

*Proof of Lemma EC.14* We prove each of (EC.46)–(EC.49) by plugging in truncated test functions into the BAR (EC.42). Using $f(z) = r_a \wedge M$, $f(z) = (q_1 + \cdots + q_n) \wedge M$ (together with the symmetry of the servers), and $f(z) = r_{s,i} \wedge M$ yields $\mathbb{E}A(1) = \lambda$, $n\mathbb{E}D_i(1) = \mathbb{E}A(1)$, and $\mathbb{E}S\mathbb{E}D_i(1) = \mathbb{P}(Q_i > 0)$, respectively, which proves (EC.46). Using $f(z) = r_a^m \wedge M$ yields the first equality in (EC.47) and using $f(z) = r_{s,i}^m \wedge M$ yields $m\mathbb{E}(R_{s,i}^{m-1}1(Q_i > 0)) = \mathbb{E}S^m\mathbb{E}D(1) = (\lambda/n)\mathbb{E}S^m$. The third equality in (EC.47) follows once we note that $\mathbb{E}(R_{s,i}^m 1(Q_i = 0)) = \mathbb{E}S^m\mathbb{P}(Q_i = 0)$ because when $Q_i(t) = 0$, then $R_{s,i}(t)$ equals the service time of the next customer to arrive. Using $f(z) = (r_a \wedge M)(r_{s,i} \wedge M)$ yields

$$\begin{aligned} &\mathbb{E}\int_0^1 R_a(t)\big(1(Q_i(t) = 0)\Lambda_i(t) + S_i(t)\big)dD_i(t) + \mathbb{E}\int_0^1 R_{s,i}(t)U(t)dA(t) \\ &= \mathbb{E}R_{s,i} + \mathbb{E}1(Q_i > 0)R_a, \end{aligned}$$

from which (EC.48) follows. The inequality (EC.49) is proved similarly using $f(z) = (r_{s,i} \wedge M)(r_{s,j} \wedge M)$. □

The second lemma is the Palm inversion formula, analogous to Lemma 3. For simplicity, we only prove it for bounded functions $f(x)$, though this assumption can be relaxed as in Lemma 3.

LEMMA EC.15. *For any bounded $f:\mathbb{R}\to\mathbb{R}$,*

$$\mathbb{E}f(X)=\mathbb{E}\Big(\int_0^1\int_0^{U(t)} f(X(t+u))du dA(t)\Big) \quad \textit{and} \tag{EC.50}$$

$$\mathbb{E}f(X)=\mathbb{E}\Big(\int_0^1\int_0^{1(Q_i(t)=0)\Lambda_i(t)+S_i(t)} f(X(t+u))du dD_i(t)\Big). \tag{EC.51}$$

*Proof of Lemma EC.15* Since $\mathbb{E}R_a<\infty$ (Lemma EC.14), the proof of (EC.50) is identical to that of (10) in Lemma 3. Similarly, to prove (EC.51), let $R_{d,i}(t)=1(Q_i(t)=0)\Lambda_i(t)+R_{s,i}(t)$ be the remaining time until the next departure from server $i$ and let $R_{d,i}$ denote the steady-state version. We only need to argue that $\mathbb{E}R_{d,i}<\infty$. This is implied by the facts that 1) the expected time until the next customer arrives to server $i$ is finite because $\mathbb{E}R_a<\infty$ and because of our uniform tie-breaking rule (meaning that an arrival will be allocated an idling server with probability at least $1/n$) and 2) $\mathbb{E}R_{s,i}<\infty$ (Lemma EC.14). □

*Proof of Proposition EC.2* The proof has two parts. In part one we assume that (EC.43)–(EC.45) hold with

$$\begin{aligned}\epsilon_0(f)=&-n\mu\delta^3\mathbb{E}\Big(\Big(-\lambda R_a+\sum_{i=1}^n\mu R_{s,i}\Big)f''(\xi)\Big)\\&+\delta^2\mu\sum_{i=1}^n\mathbb{E}\Big(1(Q_i=0)\Big(\sum_{j=1}^nQ_j-\lambda R_a+\sum_{j=1}^n\mu R_{s,j}\Big)f''(\xi)\Big),\end{aligned} \tag{EC.52}$$

$$\begin{aligned}\epsilon_A(f)=&\frac{1}{6}\delta^3\mathbb{E}\int_0^1(1-\lambda U(t))^3f'''(\xi(t))dA(t)\\&+\frac{1}{2}\delta^3c_U^2\mathbb{E}\int_0^1\Big(\sum_{i=1}^n\mu R_{s,i}(t)\Big)f'''(\xi(t-))dA(t)\\&-\frac{1}{2}\delta^2\lambda c_U^2\mathbb{E}\int_0^1\int_0^{U(t)}\big(X(t+u)-X(t-)\big)f'''(\xi(t+u))du dA(t),\end{aligned} \tag{EC.53}$$

and

$$\begin{aligned}\epsilon_{D,i}(f)=&\frac{1}{6}\delta^3\mathbb{E}\int_0^1(1-\mu S_i(t))^3f'''(\xi(t))dD_i(t)\\&+\frac{1}{2}\delta^3c_S^2\mathbb{E}\int_0^1\Big(-\lambda R_a(t)+\sum_{j\neq i}\mu R_{s,j}(t)\Big)f'''(\xi(t-))dD_i(t)\\&-\frac{1}{2}\delta^2\mu c_S^2\mathbb{E}\int_0^1\int_0^{1(Q_i(t)=0)\Lambda_i(t)}f''(X(t+u))du dD_i(t)\end{aligned}$$

$$-\frac{1}{2}\delta^2\mu c_S^2\mathbb{E}\int_0^1\int_{1(Q_i(t)=0)\Lambda_i(t)}^{1(Q_i(t)=0)\Lambda_i(t)+S_i(t)}\big(X(t+u)-X(t-)\big)f'''(\xi(t+u))dudD_i(t), \tag{EC.54}$$

and we bound the terms on the right-hand side, with the exception of the boundary term in $\epsilon_0(f)$. In part two, we expand all terms in the BAR (EC.42) to verify that (EC.43)–(EC.45) does indeed hold with $\epsilon_0(f), \epsilon_A(f)$, and $\epsilon_{D,i}(f)$ as in (EC.52)–(EC.54).

**Part one.** To prove the bound on $\epsilon_0(f)$, we bound the non-boundary term by noting that

$$n\mu\delta^3\mathbb{E}\Big(\Big(\lambda R_a+\sum_{i=1}^n \mu R_{s,i}\Big)\|f''\|\Big)\le n\mu\delta^3\|f''\|\big(\lambda^2\mathbb{E}U^2/2+(\lambda\mu\mathbb{E}S^2/2+\mu\delta\mathbb{E}S)\big),$$

where the inequality follows from Lemma EC.14. Next we bound the three terms composing $\epsilon_A(f)$. For the first term, noting that $\mathbb{E}A(1)=\lambda$, by Lemma EC.14, yields

$$\frac{1}{6}\delta^3\mathbb{E}\int_0^1(1-\lambda U(t))^3 f'''(\xi(t))dA(t)\le\frac{1}{6}\delta^3\|f'''\|\lambda\mathbb{E}\left|1-\lambda U\right|^3.$$

For the second term, we use Lemma EC.14 to get

$$\begin{aligned}
&\frac{1}{2}\delta^3 c_U^2\mathbb{E}\int_0^1\Big(\sum_{i=1}^n\mu R_{s,i}(t)\Big)f'''(\xi(t-))dA(t)\\
\le\ &\frac{1}{2}\delta^3 c_U^2\|f'''\|n\mu\lambda(\mathbb{E}R_{s,i}+\mathbb{E}R_a)\\
=\ &\frac{1}{2}\delta^3 c_U^2\|f'''\|n\mu\lambda(\rho(\lambda/n)\mathbb{E}S^2/2+\delta\mathbb{E}S+\lambda\mathbb{E}U^2/2).
\end{aligned}$$

For the third term, note that

$$\begin{aligned}
&\frac{1}{2}\delta^2\lambda c_U^2\mathbb{E}\int_0^1\int_0^{U(t)}\big(X(t+u)-X(t-)\big)f'''(\xi(t+u))dudA(t)\\
\le\ &\frac{1}{2}\delta^2\lambda c_U^2\|f'''\|\mathbb{E}\int_0^1 U(t)\sup_{0\le u\le U(t)}\left|X(t+u)-X(t-)\right|dA(t)\\
\le\ &\frac{1}{2}\delta^3\lambda c_U^2\|f'''\|(1+n+\lambda\mu\mathbb{E}U^2+\mu^2\mathbb{E}S^2).
\end{aligned}$$

To justify the second inequality, let $D_i[t,t+u]$ be the number of service completions by server $i$ on $[t,t+u]$ and note that

$$\sup_{0\le u\le U(t)}\left|X(t+u)-X(t-)\right|\le\delta+\delta\sum_{i=1}^n D_i[t,t+U(t)].$$

By forcing server $i$ to continue working even if it has no customers in its buffer, we can construct a zero-delayed counting process $\{\bar{D}_i(t), t \geq 0\}$ keeping track of *potential* service completions, which would be independent of the arrival process, and satisfy

$$D_i[t, t+u] \leq 1 + \bar{D}_i(u).$$

Since $\mathbb{E}\bar{D}_i(u) \leq \mu u + \mu^2 \mathbb{E}S^2$ by Lorden's inequality (Lorden (1970)), it follows that

$$\begin{aligned}
\mathbb{E}\int_0^1 U(t) \sup_{0\leq u\leq U(t)} |X(t+u) - X(t-)|\, dA(t) \leq& \delta\mathbb{E}\Big(U\big(1 + n + \sum_{i=1}^n \bar{D}_i(U)\big)\Big)\mathbb{E}A(1) \\
\leq& \ \delta\mathbb{E}\Big(U\big(1 + n + \mu U + \mu^2 \mathbb{E}S^2\big)\Big)\mathbb{E}A(1) \\
=& \ \delta(1 + n + \lambda\mu\mathbb{E}U^2 + \mu^2\mathbb{E}S^2) \qquad \text{(EC.55)}
\end{aligned}$$

where in the last equality we used $\mathbb{E}U\mathbb{E}A(1) = 1$. Lastly, we bound the four terms composing $\epsilon_{D,i}(f)$. First, since $\mathbb{E}D_i(1) = \lambda/n$ by Lemma EC.14, we have

$$\frac{1}{6}\delta^3\mathbb{E}\int_0^1 (1 - \mu S_i(t))^3 f'''(\xi(t))dD_i(t) \leq \frac{1}{6}\delta^3\|f'''\|\frac{\lambda}{n}\mathbb{E}\,|1 - \mu S|^3.$$

Second, using Lemma EC.14, we have

$$\begin{aligned}
&\frac{1}{2}\delta^3 c_S^2 \mathbb{E}\int_0^1 \Big(\lambda R_a(t) + \sum_{j\neq i}\mu R_{s,j}(t)\Big) f'''(\xi(t-))dD_i(t) \\
\leq& \ \frac{1}{2}\delta^3 c_S^2\|f'''\|\Big(\lambda\mu(\mathbb{E}R_{s,i} + \mathbb{E}R_a) + \sum_{j\neq i}\mu^2(\mathbb{E}R_{s,i} + \mathbb{E}R_{s,j})\Big) \\
\leq& \ \frac{1}{2}\delta^3 c_S^2\|f'''\|\Big(\lambda\mu(\rho(\lambda/n)\mathbb{E}S^2/2 + \delta\mathbb{E}S + \lambda\mathbb{E}U^2/2) + 2(n-1)\mu^2(\rho(\lambda/n)\mathbb{E}S^2/2 + \delta\mathbb{E}S)\Big) \\
=& \ \frac{1}{2}\delta^3 c_S^2\|f'''\|\Big(\frac{\lambda^2\mu\rho + 2(n-1)\mu^2\rho\lambda}{2n}\mathbb{E}S^2 + \delta(\lambda\mu + 2(n-1)\mu^2)\mathbb{E}S + \frac{1}{2}\lambda^2\mu\mathbb{E}U^2\Big).
\end{aligned}$$

Third, the Palm inversion formula (Lemma EC.15) applied to $f(z) = 1(q_i = 0)$, together with Lemma EC.14 yield

$$\mathbb{E}\int_0^1 1(Q_i(t) = 0)\Lambda_i(t)dD_i(t) = \mathbb{P}(Q_i = 0) = \delta,$$

so that

$$\frac{1}{2}\delta^2\mu c_S^2\mathbb{E}\int_0^1\int_0^{1(Q_i(t)=0)\Lambda_i(t)} f''(X(t+u))du dD_i(t) \leq \frac{1}{2}\delta^3\|f''\|c_S^2\mu\lambda/n.$$

Fourth,

$$\begin{aligned}
&\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\mathbb{E}\int_0^1\int_{1(Q_i(t)=0)\Lambda_i(t)}^{1(Q_i(t)=0)\Lambda_i(t)+S_i(t)}|X(t+u)-X(t-)|\,dudD_i(t)\\
\le{}&\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\mathbb{E}\int_0^1 S_i(t)\sup_{0\le u\le S_i(t)}|X(t+1(Q_i(t)=0)\Lambda_i(t)+u)-X(t-)|\,dD_i(t)
\end{aligned}$$

Observe that

$$\begin{aligned}
\sup_{0\le u\le S_i(t)}|X(t+u)-X(t-)|\le{}&\delta A\big[t,t+1(Q_i(t)=0)\Lambda_i(t)\big]\\
&+\delta A\big(t+1(Q_i(t)=0)\Lambda_i(t),t+1(Q_i(t)=0)\Lambda_i(t)+S_i(t)\big]\\
&+\delta\sum_{j\neq i}D_j\big[t,t+1(Q_i(t)=0)\Lambda_i(t)+S_i(t)\big],
\end{aligned}$$

where $A[a,b]$ and $D_j[a,b]$ are the number of arrivals and server $j$ departures, respectively, on $[a,b]$. To complete the bound on the fourth term in $\epsilon_{D,i}(f)$, we now bound

$$\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\mathbb{E}\int_0^1 S_i(t)\delta A\big[t,t+1(Q_i(t)=0)\Lambda_i(t)\big]dD_i(t) \tag{EC.56}$$

$$+\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\mathbb{E}\int_0^1 S_i(t)\delta A\big(t+1(Q_i(t)=0)\Lambda_i(t),t+1(Q_i(t)=0)\Lambda_i(t)+S_i(t)\big]dD_i(t) \tag{EC.57}$$

$$+\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\sum_{j\neq i}\mathbb{E}\int_0^1 S_i(t)\delta D_j\big[t,t+1(Q_i(t)=0)\Lambda_i(t)+S_i(t)\big]dD_i(t). \tag{EC.58}$$

First, we bound $A\big[t,t+1(Q_i(t)=0)\Lambda_i(t)\big]$, the number of arrivals before server $i$ gets a customer. Recall that our tie-breaking routing rule is to select among the least loaded servers uniformly at random. If server $i$ becomes idle at time $t$, then an arrival will be routed to this server with probability at least $1/n$, and it is therefore possible to construct a geometrically distributed random variable $\Gamma\ge 1$ with mean $n$ (success probability $1/n$) that a) upper bounds the number arrivals needed to route a customer to server $i$ and b) is independent of $S_i(t)$ and $\{D_j(t),t\ge 0\}_{j=1}^n$. It follows that $A\big[t,t+1(Q_i(t)=0)\Lambda_i(t)\big]\le\Gamma$. Furthermore, we define the time until the $\Gamma$th arrival as

$$\overline{\Lambda}_i(t)=R_a(t)+\sum_{m=2}^{\Gamma}U(\tau_m),$$

where $\tau_m$ is the time of the $m$th arrival after time $t$, and note that

$$1(Q_i(t)=0)\Lambda_i(t) \le \overline{\Lambda}_i(t).$$

We now have the tools needed to bound (EC.56)–(EC.58). First, we observe that

$$\begin{aligned}
&\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\mathbb{E}\int_0^1 S_i(t)\delta A\big[t, t+1(Q_i(t)=0)\Lambda_i(t)\big]dD_i(t)\\
\le{}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|\mathbb{E}\int_0^1 S_i(t)\Gamma dD_i(t) = \frac{1}{2}\delta^3\mu c_S^2\|f'''\|\mathbb{E}S\mathbb{E}\Gamma\mathbb{E}D_i(1) = \frac{1}{2}\delta^3\mu c_S^2\|f'''\|n\rho,
\end{aligned}$$

where in the last equality we used $\mathbb{E}\Gamma = n$, $\mathbb{E}S = 1/\mu$, and $\mathbb{E}D_i(1) = \lambda/n$. Second, we argue using Lorden's inequality, as we did in (EC.55), that

$$\begin{aligned}
&\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\mathbb{E}\int_0^1 S_i(t)\delta A\big(t+1(Q_i(t)=0)\Lambda_i(t), t+1(Q_i(t)=0)\Lambda_i(t)+S_i(t)\big]dD_i(t)\\
\le{}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|\mathbb{E}\int_0^1 S(1+\lambda S+\lambda^2\mathbb{E}U^2)dD_i(t)\\
={}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|\mathbb{E}\big(S(1+\lambda S+\lambda^2\mathbb{E}U^2)\big)\mathbb{E}D_i(1)\\
={}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|(\lambda/n)\big(\mathbb{E}S+\lambda\mathbb{E}S^2+\lambda^2\mathbb{E}U^2\mathbb{E}S\big).
\end{aligned}$$

Third, we bound (EC.58)

$$\begin{aligned}
&\frac{1}{2}\delta^2\mu c_S^2\|f'''\|\sum_{j\neq i}\mathbb{E}\int_0^1 S_i(t)\delta D_j\big[t, t+1(Q_i(t)=0)\Lambda_i(t)+S_i(t)\big]dD_i(t)\\
\le{}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|\sum_{j\neq i}\mathbb{E}\int_0^1 SD_j\big[t, t+\overline{\Lambda}_i+S\big]dD_i(t)\\
\le{}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|(n-1)\mathbb{E}\big(S\big(1+\mu(\overline{\Lambda}_i+S)+\mu^2\mathbb{E}S^2\big)\big)\mathbb{E}D_i(1)\\
={}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|(n-1)\big(\mathbb{E}S+\mathbb{E}\overline{\Lambda}_i+2\mu\mathbb{E}S^2\big)\lambda/n\\
\le{}& \frac{1}{2}\delta^3\mu c_S^2\|f'''\|\lambda\big(\mathbb{E}S+(\mathbb{E}R_a+\mathbb{E}U\mathbb{E}\Gamma)+2\mu\mathbb{E}S^2\big)\\
={}& \frac{1}{2}\delta^3 c_S^2\|f'''\|\lambda\big(1+(\lambda\mu\mathbb{E}U^2/2+1/\rho)+2\mu^2\mathbb{E}S^2\big).
\end{aligned}$$

**Part two.** We first prove (EC.43):

$$\delta\mathbb{E}\Big((\lambda-\sum_{i=1}^n 1(Q_i>0)\mu)f'(\widetilde{X})\Big) = -n\mu\delta^2\mathbb{E}f'(\widetilde{X})+\delta\mu\sum_{i=1}^n\mathbb{E}\big(1(Q_i=0)f'(\widetilde{X})\big).$$

Recalling that $\widetilde{X} = \delta \sum_{i=1}^n Q_i - \delta\lambda R_a + \delta \sum_{i=1}^n \mu R_{s,i}$, we have

$$-n\mu\delta^2 \mathbb{E} f'(\widetilde{X}) = -n\mu\delta^2 \mathbb{E} f'(X) - n\mu\delta^3 \mathbb{E}\Big(\Big( -\lambda R_a + \sum_{i=1}^n \mu R_{s,i}\Big) f''(\xi)\Big),$$

and

$$\begin{aligned}
&\delta\mu \sum_{i=1}^n \mathbb{E}\big(1(Q_i = 0) f'(\widetilde{X})\big) \\
&= \delta\mu \sum_{i=1}^n \mathbb{E}\Big(1(Q_i = 0)\big(f'(0) + \delta\Big(\sum_{i=1}^n Q_i - \lambda R_a + \sum_{i=1}^n \mu R_{s,i}\Big) f''(\xi)\big)\Big),
\end{aligned}$$

which proves (EC.52). The proof of (EC.53) is very similar to the proof of (9) in Proposition 1: if an arrival happens at time $t$, then $\Delta \widetilde{X}(t-) = \delta(1 - \lambda U(t))$ and, since $\mathbb{E}(1 - \lambda U) = 0$,

$$\begin{aligned}
\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-)) dA(t) = {}& \frac{1}{2}\delta^2 \mathbb{E}(1-\lambda U)^2 \mathbb{E}\int_0^1 f''(\widetilde{X}(t-)) dA(t) \\
&+ \frac{1}{6}\delta^3 \mathbb{E}\int_0^1 (1 - \lambda U(t))^3 f'''(\xi(t)) dA(t),
\end{aligned} \tag{EC.59}$$

and the first term expands into

$$\begin{aligned}
&\frac{1}{2}\delta^2 \mathbb{E}(1-\lambda U)^2 \mathbb{E}\int_0^1 f''(X(t-)) dA(t) \\
&+ \frac{1}{2}\delta^3 \mathbb{E}(1-\lambda U)^2 \mathbb{E}\int_0^1 \big(\widetilde{X}(t-) - X(t-)\big) f'''(\xi(t-)) dA(t).
\end{aligned}$$

Noting that $\mathbb{E}(1-\lambda U)^2 = c_U^2$ and $\widetilde{X}(t-) - X(t-) = \sum_{i=1}^n \mu R_{s,i}(t)$, to conclude (EC.53) we use Lemma EC.15 and repeat the argument following (11) to get

$$\begin{aligned}
&\mathbb{E}\int_0^1 f''(X(t-)) dA(t) \\
&= \lambda \mathbb{E} f''(X) - \lambda \mathbb{E}\int_0^1 \int_0^{U(t)} \big(X(t+u) - X(t-)\big) f'''(\xi(t+u)) du dA(t).
\end{aligned}$$

Lastly we prove (EC.54). Similar to (EC.59) and the display that follows, we expand the departure jump term to get

$$\mathbb{E}\int_0^1 \Delta f(\widetilde{X}(t-)) dD_i(t) = \frac{1}{2}\delta^2 \mathbb{E}(1-\mu S)^2 \mathbb{E}\int_0^1 f''(X(t-)) dD_i(t)$$

$$
\begin{aligned}
&+\frac{1}{2}\delta^3\mathbb{E}(1-\mu S)^2\mathbb{E}\int_0^1 \big(\widetilde{X}(t-)-X(t-)\big)f'''(\xi(t-))dD_i(t) \\
&+\frac{1}{6}\delta^3\mathbb{E}\int_0^1 (1-\mu S_i(t))^3 f'''(\xi(t))dD_i(t),
\end{aligned}
$$

where $\widetilde{X}(t-) - X(t-) = -\delta\lambda R_a(t) + \sum_{j\neq i}\delta\mu R_{s,j}(t)$. We conclude the proof by using Lemma EC.15 to get

$$
\begin{aligned}
\mathbb{E}f''(X) = {}& \mathbb{E}\int_0^1\int_0^{1(Q_i(t)=0)\Lambda_i(t)} f''(X(t+u))dudD_i(t) + \mathbb{E}S\int_0^1 f''(X(t-))dD_i(t) \\
&+\mathbb{E}\int_0^1\int_{1(Q_i(t)=0)\Lambda_i(t)}^{1(Q_i(t)=0)\Lambda_i(t)+S_i(t)} \big(X(t+u)-X(t-)\big)f'''(\xi(t+u))dudD_i(t).
\end{aligned}
$$

$\square$

## References

Asmussen S (2003) *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)* (New York: Springer-Verlag), second edition, ISBN 0-387-00211-1, Stochastic Modelling and Applied Probability.

Baccelli F, Brémaud P (2003) *Elements of queueing theory: Palm martingale calculus and stochastic recurrences*, volume 26 of *Applications of Mathematics* (Berlin: Springer), 2nd edition.

Barbour A (1990) Stein's method for diffusion approximations. *Probab. Theory and Related Fields* 84(3):297–322, ISSN 0178-8051, URL http://dx.doi.org/10.1007/BF01197887.

Barbour A, Ross N, Zheng G (2023) Stein's method, gaussian processes and palm measures, with applications to queueing. *The Annals of Applied Probability* 33(5):3835–3871.

Barbour AD (1988) Stein's method and Poisson process convergence. *Journal of Appl. Probab.* 25:175–184, ISSN 00219002, URL http://www.jstor.org/stable/3214155.

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research* 58(5):1398–1413, URL http://dx.doi.org/10.1287/opre.1100.0815.

Bertsimas D, Gamarnik D (2022) *Queueing theory: classical and modern methods* (Belmont, MA: Dynamic Ideas LLC), 1st edition.

Besançon E, Decreusefond L, Moyal P (2020) Stein's method for diffusive limits of queueing processes. *Queueing Systems* 95:173–201.

Blanchet J, Glynn P (2006) Complete corrected diffusion approximations for the maximum of a random walk. *Annals of Applied Probability* 16(2):951–983.

Boon M, Janssen A, van Leeuwaarden J (2023) Heavy-traffic single-server queues and the transform method. *Indagationes Mathematicae* 34(5):1014–1037.

Bramson M (2011) Stability of join the shortest queue networks. *Ann. Appl. Probab.* 21(4):1568–1625, URL http://dx.doi.org/10.1214/10-AAP726.

Braverman A (2022) The prelimit generator comparison approach of stein's method. *Stochastic Systems* 12(2):181–204, URL http://dx.doi.org/10.1287/stsy.2021.0085.

Braverman A, Dai J, Miyazawa M (2017) Heavy traffic approximation for the stationary distribution of a generalized Jackson network: the BAR approach. *Stochastic Systems* 7(1):143–196, URL http://projecteuclid.org/euclid.ssy/1495785619.

Braverman A, Dai J, Miyazawa M (2024) The BAR approach for multiclass queueing networks with SBP service policies. *Stochastic Systems* .

Braverman A, Dai JG (2017) Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems. *Ann. of Appl. Probab.* 27(1):550–581, ISSN 1050-5164, URL http://dx.doi.org/10.1214/16-AAP1211.

Braverman A, Dai JG, Feng J (2016) Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.* 6:301–366, URL http://www.i-journals.org/ssy/viewarticle.php?id=212&layout=abstract.

Brown TC, Xia A (2001) Stein's method and birth-death processes. *Ann. Probab.* 29(3):1373–1403, URL http://dx.doi.org/10.1214/aop/1015345606.

Chen LHY, Goldstein L, Shao QM (2011) *Normal approximation by Stein's method*. Probability and its Applications (New York) (Springer, Heidelberg), ISBN 978-3-642-15006-7, URL http://dx.doi.org/10.1007/978-3-642-15007-4.

Chen Y, Whitt W (2020) Extremal models for the gi/ gi/ k waiting-time tail-probability decay rate. *Operations Research Letters* 48(6):770–776.

Chen Y, Whitt W (2021) Extremal gi/gi/1 queues given two moments: exploiting tchebycheff systems. *Queueing Systems* 97(1):101–124.

Chen Y, Whitt W (2022a) Correction to: Extremal gi/gi/1 queues given two moments: exploiting tchebycheff systems. *Queueing Systems* 102(3):553–556.

Chen Y, Whitt W (2022b) Set-valued performance approximations for the queue given partial information. *Probability in the Engineering and Informational Sciences* 36(2):378–400.

Dai JG, Glynn P, Xu Y (2025) Asymptotic product-form steady-state for generalized jackson networks in multi-scale heavy traffic. URL https://arxiv.org/abs/2304.01499.

Dai JG, Guang J, Xu Y (2024) Steady-state convergence of the continuous-time jsq system with general distributions in heavy traffic. *SIGMETRICS Perform. Eval. Rev.* 52(2):39–41, ISSN 0163-5999, URL http://dx.doi.org/10.1145/3695411.3695426.

Dai JG, Hasenbein JJ, Vande Vate JH (2004) Stability and instability of a two-station queueing network. *Annals of Applied Probability* 14:326–377, ISSN 1050-5164.

Dai JG, Meyn SP (1995) Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control* 40:1889–1904.

Dai JG, Xu Y (2024) Explicit steady-state approximations for parallel server systems with heterogeneous servers. URL https://arxiv.org/abs/2406.04203.

Daley D, Kreinin AY, Trengove C (1992) Inequalities concerning the waiting time in single-server queues: A survey. *Oxford Statistical Science Series* 1(9):177–177.

Davis MHA (1984) Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of Royal Statist. Soc. series B* 46:353–388.

Eryilmaz A, Srikant R (2012) Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* 72(3-4):311–359, ISSN 0257-0130, URL http://dx.doi.org/10.1007/s11134-012-9305-y.

Feng J, Shi P (2018) Steady-state diffusion approximations for discrete-time queue in hospital inpatient flow management. *Naval Research Logistics (NRL)* 65(1):26–65, URL http://dx.doi.org/10.1002/nav.21787.

Gast N (2017) Expected values estimated via mean-field approximation are 1/n-accurate. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1(1):1–26.

Gaunt RE, Walton N (2020) Stein's method for the single server queue in heavy traffic. *Statistics & Probability Letters* 156:108566, ISSN 0167-7152, URL http://dx.doi.org/https://doi.org/10.1016/j.spl.2019.108566.

Götze F (1991) On the rate of convergence in the multivariate CLT. *Ann. Probab.* 19(2):724–739, URL http://dx.doi.org/10.1214/aop/1176990448.

Gurvich I (2014) Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *Ann. Appl. Probab.* 24(6):2527–2559, URL http://dx.doi.org/10.1214/13-AAP984.

Gurvich I, Huang J, Mandelbaum A (2014) Excursion-based universal approximations for the Erlang-A queue in steady-state. *Math. Oper. Res.* 39(2):325–373, URL http://dx.doi.org/10.1287/moor.2013.0606.

Gut A (1974) On the moments and limit distributions of some first passage times. *The Annals of Probability* 277–308.

Harrison JM, Reiman MI (1981) Reflected Brownian motion on an orthant. *Ann. Probab.* 9(2):302–308, ISSN 0091-1798, URL http://links.jstor.org/sici?sici=0091-1798(198104)9:2<302:RBMOAO>2.0.CO;2-P&origin=MSN.

Huang J, Gurvich I (2018) Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Oper. Res.* 66(4):1168–1188, URL http://dx.doi.org/10.1287/opre.2017.1715.

Hurtado-Lange D, Maguluri ST (2022) A load balancing system in the many-server heavy-traffic asymptotics. *Queueing Systems* 101(3):353–391.

Hurtado Lange DA, Maguluri ST (2022) Heavy-traffic analysis of queueing systems with no complete resource pooling. *Mathematics of Operations Research* 47(4):3129–3155, URL http://dx.doi.org/10.1287/moor.2021.1248.

Kingman JFC (1961) The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society* 57:902–904, URL http://dx.doi.org/10.1017/S0305004100036094.

Kingman JFC (1962) On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B* 24:383–392, ISSN 0035-9246, URL http://links.jstor.org/sici?sici=0035-9246(1962)24:2<383:OQIHT>2.0.CO;2-A&origin=MSN.

Köllerström J (1976) Stochastic bounds for the single-server queue. *Mathematical Proceedings of the Cambridge Philosophical Society* 80(3):521–525, URL http://dx.doi.org/10.1017/S0305004100053135.

Li J, Ou J (1995) Characterizing the idle-period distribution of gi/g/1 queues. *Journal of applied probability* 32(1):247–255.

Lieberman GM (2013) *Oblique Derivative Problems for Elliptic Equations* (WORLD SCIENTIFIC), URL http://dx.doi.org/10.1142/8679.

Lorden G (1970) On excess over the boundary. *The Annals of Mathematical Statistics* 41(2):520 – 527, URL http://dx.doi.org/10.1214/aoms/1177697092.

Loulou R (1978) An explicit upper bound for the mean busy period in a $GI/G/1$ queue. *Journal of Applied Probability* 15(2):452–455, URL http://dx.doi.org/10.2307/3213419.

Mackey L, Gorham J (2016) Multivariate Stein factors for a class of strongly log-concave distributions. *Electron. Commun. Probab.* 21:14, URL http://dx.doi.org/10.1214/16-ECP15.

Maguluri ST, Burle SK, Srikant R (2018) Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *Queueing Systems* 88(3):279–309.

Maguluri ST, Srikant R (2016) Heavy traffic queue length behavior in a switch under the maxweight algorithm. *Stochastic Systems* 6(1):211–250, URL http://dx.doi.org/10.1214/15-SSY193.

Miyazawa M (2015) Diffusion approximation for stationary analysis of queues and their networks: a review. *J. Oper. Res. Soc. Japan* 58(1):104–148, ISSN 0453-4514, URL http://dx.doi.org/10.15807/jorsj.58.104.

Miyazawa M (2017) A unified approach for large queue asymptotics in a heterogeneous multiserver queue. *Adv. in Appl. Probab.* 49(1):182–220, ISSN 0001-8678, URL http://dx.doi.org/10.1017/apr.2016.84.

Peköz EA, Röllin A (2011) New rates for exponential approximation and the theorems of Rényi and Yaglom. *The Annals of Probability* 39(2):587 – 608, URL http://dx.doi.org/10.1214/10-AOP559.

Ross N (2011) Fundamentals of Stein's method. *Probab. Surv.* 8:210–293, ISSN 1549-5787, URL http://dx.doi.org/10.1214/11-PS182.

Siegmund D (1979) Corrected diffusion approximations in certain random walk problems. *Advances in Applied Probability* 11(4):701–719, ISSN 00018678, URL http://www.jstor.org/stable/1426855.

Stein C (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, 583–602 (Berkeley, Calif.: University of California Press), URL http://projecteuclid.org/euclid.bsmsp/1200514239.

Stolyar AL (2015) Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime. *Stoch. Syst.* 5(2):239–267, URL http://dx.doi.org/10.1214/14-SSY139.

Ward A, Glynn P (2003) A diffusion approximation for a markovian queue with reneging. *Queueing Systems* 43(1-2):103–128, ISSN 0257-0130, URL http://dx.doi.org/10.1023/A:1021804515162.

Whitt W (1986) Deciding which queue to join: Some counterexamples. *Oper. Res.* 34(1):55–62, ISSN 0030-364X.

Wolff RW, Wang CL (2003) Idle period approximations and bounds for the gi/g/1 queue. *Advances in Applied Probability* 35(3):773–792.

Ying L (2016) On the approximation error of mean-field models. *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, 285–297 (Antibes Juan-les-Pins, France: ACM), URL http://dx.doi.org/10.1145/2964791.2901463.

Ying L (2017) Stein's method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Meas. Anal. Comput. Syst.* 1(1):12:1–12:27, ISSN 2476-1249, URL http://dx.doi.org/10.1145/3084449.

Zhou X, Shroff N (2020) A note on stein's method for heavy-traffic analysis. *arXiv preprint arXiv:2003.06454* .