Queueing Systems 46, 507–536, 2004

Heavy-Traffic Limits for Loss Proportions in Single-Server Queues

WARD WHITT ward.whitt@columbia.edu
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA

Received 1 August 2002; Revised 10 December 2002

Abstract. We establish heavy-traffic stochastic-process limits for the queue-length and overflow stochastic processes in the standard single-server queue with finite waiting room ($G/G/1/K$). We show that, under regularity conditions, the content and overflow processes in related single-server models with finite waiting room, such as the finite dam, satisfy the same heavy-traffic stochastic-process limits. As a consequence, we obtain heavy-traffic limits for the proportion of customers or input lost over an initial interval. Except for an interchange of the order of two limits, we thus obtain heavy-traffic limits for the steady-state loss proportions. We justify the interchange of limits in $M/GI/1/K$ and $GI/M/1/K$ special cases of the standard $GI/GI/1/K$ model by directly establishing local heavy-traffic limits for the steady-state blocking probabilities.

Keywords: single-server queue, finite-capacity queue, finite dam, blocking probability, loss rate, overflows, heavy-traffic, local limit, diffusion approximation, heavy-tailed probability distribution, stable process

1. Introduction

One of the important lessons from the heavy-traffic limits for the queue-length and waiting-time processes in the general single-server queue ($G/G/1/K$), with either infinite or finite waiting room ($K \leqslant \infty$), is that the heavy-traffic limits depend upon the arrival process and the service times only through the scaling constants that appear in their functional central limit theorems (which are assumed to hold); e.g., see [20, chapters 5, 8 and 9] and section 2 here. For the special case of a renewal arrival process independent of a sequence of independent and identically distributed (IID) service times ($GI/GI/1/K$), where the second moments of the interarrival and service times are finite, the relevant parameters are the first two moments of the interarrival-time and service-time distributions when they exist.

A shortcoming of that heavy-traffic limit for the model with finite waiting room, even when appropriately extended to steady-state distributions (which often can be done), is that it does not directly imply an associated heavy-traffic limit for the steady-state blocking probabilities (or long-run loss proportions). That requires a refined local heavy-traffic limit. In particular, we need a limit for the associated sequence of probability mass functions of the steady-state queue length at arrival epochs (just before the arrival), evaluated at the upper boundary.

However, the steady-state loss proportions also can be approached in another way: We can directly establish a heavy-traffic limit for the loss proportions over initial intervals. Heavy traffic is achieved by considering a sequence of models indexed by n and letting $\rho_n \to 1$ and $K_n \to \infty$, where ρ_n is the traffic intensity and K_n is the number of waiting spaces in model n. To quickly see the main idea, suppose that $\Pi_n(t)$ represents the proportion of arrivals over the time interval $[0, t]$ that are lost (blocked). In considerable generality, when stochastic-process limits hold with time scaling by n and space scaling by c_n, where $c_n \to \infty$ and $n/c_n \to \infty$ with

$$\frac{n(1-\rho_n)}{c_n} \to \eta \quad \text{for } -\infty < \eta < \infty \tag{1.1}$$

and

$$\frac{K_n}{c_n} \to \kappa \quad \text{for } 0 < \kappa < \infty, \tag{1.2}$$

it is possible to show that

$$\frac{n}{c_n}\Pi_n(nt) \Rightarrow \Pi(t) \quad \text{as } n \to \infty, \tag{1.3}$$

where $\Rightarrow$ denotes convergence in distribution and $\Pi(t)$ is the loss proportion associated with a limiting stochastic process, such as a reflected Brownian motion (RBM), which is defined by the upper-regulator map associated with the two-sided reflection map.

Assuming that $\Pi(t) \Rightarrow \pi$ as $t \to \infty$, where π is a deterministic steady-state loss proportion, we obtain the associated iterated double limit

$$\frac{n}{c_n}\Pi_n(nt) \Rightarrow \pi \tag{1.4}$$

by *first* letting $n \to \infty$ to get (1.3) and *second* letting $t \to \infty$.

Assuming that *both* $\Pi_n(nt) \Rightarrow \pi_n$ as $t \to \infty$ for each n and $\Pi(t) \Rightarrow \pi$ as $t \to \infty$, where π_n and π are deterministic steady-state loss proportions, we can obtain the associated limit

$$\lim_{n\to\infty} \frac{n}{c_n}\pi_n \to \pi \tag{1.5}$$

from (1.4) if we can interchange the order of the two limits.

Based on the iterated double limit in (1.4) (hoping that the interchange required for (1.5) is valid), we thus generate the candidate heavy-traffic approximation

$$\pi_n \approx \frac{c_n}{n}\pi, \tag{1.6}$$

which should be very useful because π usually is much easier to compute than π_n. Moreover, π typically depends on only limited information about the original model.

In this paper we elaborate on this approach to heavy-traffic approximations for steady-state loss proportions or blocking probabilities. We establish a heavy-traffic stochastic-process limit for the general $G/G/1/K$ queue supporting (1.3) and we investigate when interchanging the two limits in (1.4) is valid. To support the validity of interchanging the order of the two limits, we establish local heavy-traffic limits for the blocking probability in the $M/GI/1/K$ model when the sequence of traffic intensities approach 1 from below and in the $GI/M/1/K$ model when the sequence of traffic intensities approach 1 from above. In these cases, we require that the general distribution for the service time or interarrival time have a finite second moment, so we are in the domain of heavy-traffic diffusion approximations. We show that the local heavy-traffic limits are fully consistent with the heuristic diffusion approximation in (1.6) based on the limit in (1.4).

We are aware of few related heavy-traffic local limits. The local heavy-traffic limit for the blocking probability in the Markovian $M/M/1/K$ model is relatively elementary, and was established by Berger and Whitt [1]. The local heavy-traffic limit for the blocking probability in the $GI/M/K/0$ multiserver pure-loss model was established by Borovkov [4]; see theorem 15(2) on p. 226. That local heavy-traffic limit is discussed in [18]. These two previous local heavy-traffic limits are consistent with the heuristic diffusion approximation in (1.6). Establishing other local heavy-traffic limits remains an important direction of research.

The heuristic diffusion approximation in (1.6) is applied in [21] to develop approximations for the steady-state blocking probability in the $G/GI/n/K$ model. The local heavy-traffic limit here for the $GI/M/1/K$ model can be combined with the heavy-traffic stochastic process limit for the $G/M/n/K$ model in [22] to show that the heuristic approximation in (1.6) is asymptotically correct in the special case of the $GI/M/n/K$ model in which the traffic intensities approach 1 from above. The reason is that the $GI/M/n/K$ model behaves like an associated $GI/M/1/K$ model when all servers are busy. Thus, this paper verifies the conjecture about supporting local heavy-traffic limits in [21, remark 7.1] in one special case.

Here is how the rest of this paper is organized: In section 2 we establish a general version of the heavy-traffic stochastic-process limit for the $G/G/1/K$ queue, allowing for limit processes with discontinuous sample paths. In section 3 we elaborate on the $G/GI/1/K$ special case in which the service-time distribution has a heavy tail. Then the scaled queue-length processes converge to a reflected stable Lévy motion. For this case, we calculate the steady-state loss proportion for the limit process.

Then we establish the local heavy-traffic limits for the $M/GI/1/K$ and $GI/M/1/K$ models in sections 4 and 5. In section 6 we discuss corresponding heavy-traffic limits and approximations for related finite-capacity single-server queueing models, such as the finite dam. When appropriate limit processes have continuous sample paths, we show that the same heavy-traffic stochastic-process limits hold for these other models, showing that these different models are asymptotically equivalent in heavy-traffic (with the heavy-traffic scaling in which the finite capacities are allowed to grow), even though they can have quite different performance under other circumstances.

We postpone several proofs until section 7. We draw conclusions in section 8. For more on approximations for finite-capacity single-server queues, see [17, section 4.3]. For alternative approximations based on limits in which only $K \to \infty$, with the traffic intensity remaining fixed, see [12,23].

2. A stochastic-process limit for the $G/G/1/K$ queue

In this section we establish a heavy-traffic stochastic-process limit for the queue-length and loss (overflow) stochastic processes in the standard $G/G/1/K$ model, allowing general scaling. The result is similar to the limit for the fluid model in [20, section 5.4], but more difficult because the two-sided reflection map cannot be applied directly. The result here is a modification of the corresponding infinite-capacity limit in [20, theorem 9.3.4], filling a gap in [20]. The first heavy-traffic stochastic-process limit for a single-server queue with finite waiting room was established by Kennedy [13].

We consider a sequence of models indexed by n with heavy-traffic scaling. Just as in [20, section 9.3], for each n there is a sequence of ordered pairs of nonnegative random variables $\{(U_{n,k}, V_{n,k})\colon k \geqslant 1\}$. The variable $U_{n,k}$ represents the interarrival time between customers k and $k-1$, while the variable $V_{n,k}$ represents the service time of the kth customer to receive service. (Service times are not assigned to blocked customers. That is without loss of generality if the service times are IID, but could be a restriction otherwise.) For simplicity, we assume that the first customer arrives at time $U_{n,1}$ to find an empty system.

Form associated partial sums by letting

$$\begin{aligned} S^u_{n,k} &\equiv U_{n,1} + \cdots + U_{n,k}, \\ S^v_{n,k} &\equiv V_{n,1} + \cdots + V_{n,k} \end{aligned} \tag{2.1}$$

with

$$S^u_{n,0} \equiv S^v_{n,0} \equiv U_{n,0} \equiv V_{n,0} \equiv 0. \tag{2.2}$$

Let $A_n \equiv \{A_n(t)\colon t \geqslant 0\}$ and $N_n \equiv \{N_n(t)\colon t \geqslant 0\}$ be counting processes associated with the interarrival times and service times, i.e.,

$$\begin{aligned} A_n(t) &\equiv \max\bigl\{k \geqslant 0\colon S^u_{n,k} \leqslant t\bigr\}, \\ N_n(t) &\equiv \max\bigl\{k \geqslant 0\colon S^v_{n,k} \leqslant t\bigr\}, \quad t \geqslant 0. \end{aligned} \tag{2.3}$$

We will state results only for the queue-length and loss processes. (Limits for other related processes can be obtained by methods in [20].) Let $Q_n(t)$ be the queue length (number in system, including the customer in service, if any) at an arbitrary time t. Let $L_n(t)$ be the number of customers lost (blocked) in the interval [0, t].

We now form associated scaled stochastic processes: Let $\lfloor t \rfloor$ be the greatest integer less than or equal to t and let

$$\begin{aligned}
\mathbf{S}_n^u(t) &\equiv \frac{S_{n,\lfloor nt \rfloor}^u - \rho_n^{-1}\lfloor nt \rfloor}{c_n}, \\
\mathbf{S}_n^v(t) &\equiv \frac{S_{n,\lfloor nt \rfloor}^v - \lfloor nt \rfloor}{c_n}, \\
\mathbf{A}_n(t) &\equiv \frac{A_n(\lfloor nt \rfloor) - \rho_n nt}{c_n}, \\
\mathbf{N}_n(t) &\equiv \frac{N_n(\lfloor nt \rfloor) - nt}{c_n}, \\
\mathbf{Q}_n(t) &\equiv \frac{Q_n(nt)}{c_n}, \\
\mathbf{L}_n(t) &\equiv \frac{L_n(nt)}{c_n}, \quad t \geqslant 0.
\end{aligned} \tag{2.4}$$

Without loss of generality, we have chosen the time units so that the translation scaling constants for the service times in (2.4) are 1, while the translation scaling constants for the interarrival times in (2.4) are ρ_n^{-1}, where ρ_n is yet to be specified, but will usually be the traffic intensity. The canonical case is stationary sequences with mean service time 1 and mean interarrival time $1/\rho_n$.

To state the result, let $\Rightarrow$ denote convergence in distribution (weak convergence) and let $D \equiv D([0,\infty), \mathbb{R}, M_1) \equiv (D, M_1)$ denote the space of all right-continuous real-valued functions on the interval $[0,\infty)$ with left limits everywhere (except at 0), endowed with the Skorohod M_1 topology [20]. As discussed at length in [20], especially in chapter 6, the unconventional M_1 topology is needed when the limit process has discontinuous sample paths, as occurs in the standard $G/G/1/K$ model with heavy-tailed service-time distributions. At continuous limit functions, convergence in (D, M_1) corresponds to uniform convergence over bounded intervals. Let $D^k \equiv (D, M_1)^k$ denote the k-fold product of D with itself, endowed with the product topology, where the M_1 topology is used on each coordinate space. Let $Disc(x)$ be the set of discontinuities of the function x. Let $\overset{\mathrm{d}}{=}$ denote equality in distribution. Let $\mathbf{e}$ be the identity function in D, i.e., $\mathbf{e}(t) \equiv t$, $t \geqslant 0$.

Let $(\phi_0, \psi_0^L) : D \to D^2$ be the one-sided reflection map with lower boundary at 0, satisfying

$$\phi_0(x) = x + \psi_0^L(x), \tag{2.5}$$

and let $(\phi_{0,\kappa}, \psi_0^L, \psi_\kappa^U) : D \to D^3$ be the two-sided reflection map with lower boundary at 0 and upper boundary at κ, satisfying

$$\phi_{0,\kappa}(x) = x + \psi_0^L(x) - \psi_\kappa^U(x), \tag{2.6}$$

where ϕ_0 and $\phi_{0,\kappa}$ are the content functions, ψ_0^L is the lower-boundary regulator function and ψ_κ^U is the upper-boundary regulator function satisfying the usual properties; see [11, pp. 17–24], or [20, sections 5.2, 13.5 and 14.8].

We are now ready to state our main general result. As indicated above, we use the M_1 topology on the function space D. The assumed convergence in the M_1 topology in condition (2.9) below is implied by convergence in the standard J_1 topology, because the J_1 topology is stronger than the M_1 topology. However, the conclusion in (2.12) is only convergence in the M_1 topology. When the limit processes have continuous sample paths w.p.1, the M_1 and J_1 topologies both coincide with the topology of uniform convergence over bounded intervals. The proof of the following result appears in section 7 along with several others.

Theorem 2.1. Consider the sequence of $G/G/1/K$ models specified above. Suppose that $c_n \to \infty$, $n/c_n \to \infty$,

$$\eta_n \equiv \frac{n(1-\rho_n)}{c_n} \to \eta \quad \text{for } -\infty < \eta < \infty, \tag{2.7}$$

and

$$\frac{K_n}{c_n} \to \kappa \quad \text{for } 0 < \kappa < \infty \tag{2.8}$$

as $n \to \infty$. Suppose that

$$\left(\mathbf{S}_n^u, \mathbf{S}_n^v\right) \Rightarrow \left(\mathbf{S}^u, \mathbf{S}^v\right) \quad \text{in } (D, M_1)^2, \tag{2.9}$$

where

$$P\left(\mathbf{S}^u(0) = 0\right) = P\left(\mathbf{S}^v(0) = 0\right) = 1 \tag{2.10}$$

and

$$P\left(Disc\left(\mathbf{S}^u\right) \cap Disc\left(\mathbf{S}^v\right) = \phi\right) = 1. \tag{2.11}$$

Then

$$(\mathbf{Q}_n, \mathbf{L}_n) \Rightarrow (\mathbf{Q}, \mathbf{L}) \equiv \left(\phi_{0,\kappa}\left(\mathbf{S}^v - \mathbf{S}^u - \eta\mathbf{e}\right), \psi_\kappa^U\left(\mathbf{S}^v - \mathbf{S}^u - \eta\mathbf{e}\right)\right) \quad \text{in } (D, M_1)^2. \tag{2.12}$$

Theorem 2.1 implies an associated limit for the loss proportion over an initial interval. Let $\Pi_n(t)$ be the proportion of arrivals lost over the time interval $[0, t]$ in model n (with the capitalization indicating it is a random quantity). For $y \in \mathbb{R}$, let $[y]_a \equiv \max\{y, a\}$. There is some difficulty in the definition and the convergence for very small t (which is not of interest to us). Clearly, a reasonable definition is

$$\Pi_n(t) \equiv \frac{L_n(t)}{[A_n(t)]_1}, \quad t \geqslant 0. \tag{2.13}$$

(We use $[A_n(t)]_1$ in the denominator to avoid dividing by 0. We could instead directly let $\Pi_n(t) \equiv 0$ when $A_n(t) = 0$.) Let the associated scaled random function be

$$\boldsymbol{\Pi}_n(t) \equiv \frac{n\Pi_n(nt)}{c_n}, \quad t \geqslant 0. \tag{2.14}$$

Note that $\mathbf{L}(t)/t$ might not be well defined at $t = 0$. Thus we establish the limit in the space $D((0,\infty), \mathbb{R}, M_1)$. Convergence $x_n \to x$ in $D((0,\infty), \mathbb{R})$ in any of the topologies is equivalent to convergence of the restrictions in $D([a,b], \mathbb{R})$, with the same topology, for all a, b, with $0 < a < b < \infty$ that are continuity points of x.

Corollary 2.1. Under the assumptions of theorem 2.1,

$$\boldsymbol{\Pi}_n \Rightarrow \boldsymbol{\Pi} \quad \text{in } D\big((0,\infty), \mathbb{R}, M_1\big), \tag{2.15}$$

where $\boldsymbol{\Pi}(t) \equiv \mathbf{L}(t)/t$, $t \geqslant 0$. Suppose that $P(t \in Disc(\mathbf{L})) = 0$ for all $t > 0$. Then

$$\frac{n}{c_n}\Pi_n(nt) \Rightarrow \boldsymbol{\Pi}(t) \quad \text{as } n \to \infty \tag{2.16}$$

and

$$\frac{\Pi_n(nt)}{|1-\rho_n|} \Rightarrow \frac{\boldsymbol{\Pi}(t)}{|\eta|} \quad \text{as } n \to \infty \tag{2.17}$$

for all $t > 0$, where η is the limit in (2.7). If $\boldsymbol{\Pi}(t) \Rightarrow \pi$ as $t \to \infty$, then

$$\frac{n}{c_n}\Pi_n(nt) \Rightarrow \pi \tag{2.18}$$

as first $n \to \infty$ and then $t \to \infty$.

Remark 2.1. Heavy-traffic scaling. As discussed in [20, section 5.5], it is revealing to index the models by ρ instead of n and then let $\rho \to 1$. Since we have finite waiting rooms, it is natural to allow $\rho > 1$ as well as $\rho < 1$. However, what we say here does not apply to the case $\rho = 1$.

As explained in [20, section 5.5], the canonical heavy-traffic scaling of time and space as functions of ρ when $\rho \neq 1$ and $c_n = n^H$ for $0 < H < 1$ is

$$n_\rho = \left|\frac{\eta}{1-\rho}\right|^{1/(1-H)} \quad \text{and} \quad c_\rho = n_\rho^H = \left|\frac{\eta}{1-\rho}\right|^{H/(1-H)}, \tag{2.19}$$

where η comes from (2.7). (Since $\mu = 1$ here, $\zeta = \eta$ in (5.7) on p. 160 of [20].) When we index by ρ instead of by n, the limit (2.17) becomes

$$\Pi_\rho(n_\rho t) \sim \left|\frac{(1-\rho)}{\eta}\right| \Pi(t) \quad \text{as } \rho \to 1 \quad (\rho \uparrow 1 \text{ or } \rho \downarrow 1), \tag{2.20}$$

where $\sim$ means that the ratio of the two sides converges to 1 as $\rho \to 1$. We obtain the important practical insight that π_ρ should be of order $|1-\rho|$ as $\rho \to 1$, independent of the way c_n grows with n (subject to the stated conditions), provided that the number of waiting spaces grows according to (2.8).

For the heavy-traffic loss approximations, condition (2.8) specifying how the waiting room must grow clearly plays a critical role. Even though the scaling exponent H plays no direct role in (2.20), the limit (2.20) does depend on H through condition (2.8), which expresses how the waiting room grows. With indexing by ρ, the canonical waiting room size is

$$K_\rho = \kappa c_\rho = \kappa \left| \frac{\eta}{1-\rho} \right|^{H/(1-H)}. \tag{2.21}$$

Given the scaling in (2.19) and (2.21), we get (2.20).

Given that π_ρ and K_ρ are both indexed by ρ, we can go further and index π by the waiting room size K (without changing the limiting regime in which $\rho \to 1$). Combining (2.19) and (2.21), we obtain the approximation

$$\pi_K \approx \frac{b}{K^{(1-H)/H}}, \tag{2.22}$$

where $b = \pi\kappa^{(1-H)/H}$ is a constant, with κ coming from (2.8) and π coming from (2.17) and (2.18). Except possibly for the constant b (see remark 3.1), this asymptotic form obtained in the heavy-traffic regime is consistent with the asymptotic behavior as $K \to \infty$ with the traffic intensity held fixed in special cases; e.g., see [12] and [23]. (They use $H = 1/\alpha$; then $(1-H)/H = \alpha - 1 > 0$.) The general form of the approximation in (2.22) provides important practical insight, even if the constant $b = \pi\kappa^{\alpha-1}$ is not calculated. However, we go further to calculate the constant in (2.22) as well; e.g., see (2.28) for the light-tailed Brownian case and theorem 3.2 for a heavy-tailed case.

Given that we want heavy-traffic limits for the steady-state blocking probability, we have gone a long way with the double limit in (2.18). The remaining problem is to interchange the order of the two limits in (2.18). Thus, just as with heavy-traffic approximations for other steady-state quantities, (2.18) provides strong support for the approximation $\pi_n \approx (c_n/n)\pi$ in (1.6), where it is assumed that $\mathbf{L}(t)/t \Rightarrow \pi$ as $t \to \infty$ and $\Pi_n(t) \Rightarrow \pi_n$ as $t \to \infty$ for each n, with π_n and π both being deterministic. However, little is yet known about when this interchange of limits is valid. We establish positive results for two special cases in sections 4 and 5. For the general infinite-capacity $G/G/1/\infty$ model, related investigations have been carried out by Szczotka [14,15].

Remark 2.2. The standard Brownian case. The standard case has scaling by $c_n = \sqrt{n}$ with the limit processes in (2.9) being $\mathbf{S}^u = \sqrt{c_a^2}\mathbf{B}^1$ and $\mathbf{S}^v = \sqrt{c_s^2}\mathbf{B}^2$, where $\mathbf{B}^1$ and $\mathbf{B}^2$ are independent standard (zero drift, unit diffusion coefficient) Brownian motions. A special case of the standard case is the $GI/GI/1/K$ model under the condition that the first two moments of the interarrival and service times are finite. In that special case, the constants c_a^2 and c_s^2 are the squared coefficients of variation (SCV, variance divided by the square of the mean) of an interarrival time and a service time, respectively.

In the standard (Brownian) case, conditions (2.10) and (2.11) automatically hold and

$$\mathbf{S}^v - \mathbf{S}^u - \eta\mathbf{e} \overset{\mathrm{d}}{=} \sigma\mathbf{B} - \eta\mathbf{e}, \tag{2.23}$$

where **B** is a standard Brownian motion and

$$\sigma^2 = c_a^2 + c_s^2. \tag{2.24}$$

The limit process $\phi_{0,\kappa}(\mathbf{S}^v - \mathbf{S}^u - \eta\mathbf{e})$ in theorem 2.1 is reflected Brownian motion (RBM) with drift, using the two-sided reflection map. As noted above, in the $GI/GI/1/K$ special case, which includes sections 4 and 5, c_a^2 and c_s^2 are the SCVs of an interarrival time and a service time.

For RBM with parameters η and σ as in (2.23), $\mathbf{Q}(t) \Rightarrow \mathbf{Q}(\infty)$ as $t \to \infty$, where the probability distribution of $\mathbf{Q}(\infty)$ is absolutely continuous with the density

$$f_{\mathbf{Q}(\infty)}(x) = \frac{2\eta \mathrm{e}^{-2\eta x/\sigma^2}}{\sigma^2(1 - \mathrm{e}^{-2\eta\kappa/\sigma^2})}, \quad 0 \leqslant x \leqslant \kappa, \tag{2.25}$$

when $\eta \neq 0$, and the uniform density on $[0, \kappa]$ when $\eta = 0$; see [11, pp. 86–92].

Henceforth in this remark, assume that $\eta > 0$ (corresponding to $\rho_n < 1$). Let $\mathbf{Q}^\infty(\infty)$ have the steady-state distribution of $RBM(\eta, \sigma)$ with no upper barrier. The distribution of $\mathbf{Q}(\infty)$ is just the conditional distribution of $\mathbf{Q}^\infty(\infty)$ given that the upper barrier κ is not exceeded, i.e.,

$$P\big(\mathbf{Q}(\infty) \leqslant x\big) = \frac{P(\mathbf{Q}^\infty(\infty) \leqslant x)}{P(\mathbf{Q}^\infty(\infty) \leqslant \kappa,)}, \quad 0 \leqslant x \leqslant \kappa. \tag{2.26}$$

Turning to the loss proportion, by [11],

$$\frac{\mathbf{L}(t)}{t} \to \pi \equiv \frac{\sigma^2}{2} f_{\mathbf{Q}(\infty)}(\kappa) = \frac{\eta \mathrm{e}^{-2\eta\kappa/\sigma^2}}{(1 - \mathrm{e}^{-2\eta\kappa/\sigma^2})}. \tag{2.27}$$

Hence, in the standard Brownian case, the approximation (1.6) based on corollary 2.1 becomes

$$\pi_n \approx \frac{\pi}{\sqrt{n}} = \frac{\eta \mathrm{e}^{-2\eta\kappa/\sigma^2}}{\sqrt{n}(1 - \mathrm{e}^{-2\eta\kappa/\sigma^2})}. \tag{2.28}$$

As discussed in [20, section 2.4.1], a rough approximation for the loss proportion is the infinite-capacity steady-state tail probability. For $RBM(\eta, \sigma)$,

$$P\big(\mathbf{Q}^\infty(\infty) > \kappa\big) = \mathrm{e}^{-2\eta\kappa/\sigma^2}. \tag{2.29}$$

Combining (2.28) and (2.29), the two-step rough approximation is

$$\pi_n \approx \frac{\pi}{\sqrt{n}} \approx \frac{P(\mathbf{Q}^\infty(\infty) > \kappa)}{\sqrt{n}} = \frac{\mathrm{e}^{-2\eta\kappa/\sigma^2}}{\sqrt{n}}. \tag{2.30}$$

Note that (2.30) differs from (2.28) only by the factor $\eta/(1 - \exp(-2\eta\kappa/\sigma^2))$. For small loss proportions, this factor tends to be negligible compared to the exponential term. In the heavy-traffic setting we would of course use approximation (2.28) instead of (2.30). The ratio of (2.28) to (2.30) could serve as a multiplicative refinement when the steady-state tail probability $P(Q_n^\infty(\infty) > K_n)$ is calculated directly.

3. Losses with heavy-tailed service-time distributions

In this section we consider a nonstandard case in greater detail. We now consider the $G/GI/1/K$ model, in which the service times are independent of the arrival process and IID, where the service-time cumulative distribution function (cdf), say G, has a heavy tail. Specifically, we assume that

$$G^c(t) \sim \gamma t^{-\alpha} \quad \text{as } t \to \infty \text{ for } 1 < \alpha < 2, \tag{3.1}$$

where γ is some positive constant and $G^c(t) \equiv 1 - G(t)$, which implies that G has finite mean but infinite variance.

Condition (3.1) is a special case of a regulaly varying tail with index $-\alpha$; see [3] or [20, appendix A]. We do not express results in that more general framework, because we believe that the pure-power-tail case in (3.1) is more natural for applications. The scaling becomes more complicated with the more general regularly-varying tail, but it has been studied extensively; see [20, theorem 4.5.1 and (5.25)]. For the general regularly-varying case, Boxma and Cohen [5] have determined appropriate scaling for the steady-state waiting-time distributions via a solution of an equation, leading to their “coefficient of contraction.” From their (4.9), it is evident that the simple power scaling here can be used with the pure power tail in (3.1).

We state a consequence of theorem 2.1 in this $G/GI/1/K$ setting with heavy-tailed service time distribution. We obtain convergence to a reflected stable Lévy motion (RSLM); see [20, sections 4.5 and 8.5] for background. Let $S_\alpha(\sigma, \beta, \mu)$ denote the stable law (or random variable with that stable law) with index α, scale parameter σ, skewness parameter β and shift parameter μ, as in [20, section 4.5].

Theorem 3.1. Consider a sequence of $G/GI/1/K$ models specified as in section 2, where the service-time cdf G satisfies (3.1). Let the scaling constants be $c_n = n^{1/\alpha}$. Suppose that conditions (2.7) and (2.8) in theorem 2.1 hold and

$$\mathbf{S}_n^u \Rightarrow 0\mathbf{e}. \tag{3.2}$$

Then

$$(\mathbf{Q}_n, \mathbf{L}_n) \Rightarrow (\mathbf{Q}, \mathbf{L}) \equiv \big(\phi_{0,\kappa}(\mathbf{S}^v - \eta\mathbf{e}), \psi_\kappa^U(\mathbf{S}^v - \eta\mathbf{e})\big) \quad \text{in } (D, M_1)^2, \tag{3.3}$$

where $\mathbf{S}^v$ is a stable Lévy motion with

$$\mathbf{S}^v(1) \overset{\mathrm{d}}{=} \sigma S_\alpha(1, 1, 0), \tag{3.4}$$

and

$$\sigma = \left(\frac{\gamma}{c_\alpha}\right)^{1/\alpha}, \tag{3.5}$$

with γ coming from (3.1) and

$$c_\alpha \equiv \frac{\alpha - 1}{\Gamma(2-\alpha)|\cos(\pi\alpha/2)|}. \tag{3.6}$$

If, in addition, $\eta > 0$ (corresponding to $\rho_n < 1$), as $t \to \infty$, $\mathbf{Q}(t) \Rightarrow \mathbf{Q}(\infty)$, where

$$H(x) \equiv P\big(\mathbf{Q}(\infty) \leqslant x\big) = \frac{H^\infty(x)}{H^\infty(\kappa)}, \quad 0 \leqslant x \leqslant \kappa, \tag{3.7}$$

with H^∞ being the cdf of the limiting RSLM in the infinite-capacity case, which is absolutely continuous with density h^∞, implying that H is absolutely continuous with density $h = h^\infty / H^\infty(\kappa)$. The density h^∞ has Laplace transform

$$\hat{h}^\infty(s) \equiv \int_0^\infty \mathrm{e}^{-sx} h^\infty(x)\,\mathrm{d}x = \frac{1}{1 + (\xi s)^{\alpha-1}}, \tag{3.8}$$

where the scale parameter ξ ($H_\xi^\infty(x) = H_1^\infty(\xi x)$) satisfies

$$\xi^{\alpha-1} \equiv \frac{-\sigma^\alpha}{\eta\cos(\pi\alpha/2)} > 0. \tag{3.9}$$

Proof and discussion. Condition (3.1) implies that the cdf G is in the normal domain of attraction of a stable law with index α. By theorems 4.5.2, 4.5.3 and 7.3.2 and corollary 7.3.2 of [20],

$$\big(\mathbf{S}_n^v, \mathbf{N}_n\big) \Rightarrow \big(\mathbf{S}^v, -\mathbf{S}^v\big) \quad \text{in } (D, M_1)^2 \tag{3.10}$$

as $n \to \infty$ when $c_n = n^{1/\alpha}$, where $\mathbf{S}_n^v$ and $\mathbf{N}_n$ are defined as in (2.1)–(2.4) and $\mathbf{S}^v$ is a stable Lévy motion. Hence, here the general scaling discussed in remark 2.1 holds with $H = 1/\alpha$. Theorem 4.5.2 of [20] implies that the scale parameter σ is as given in (3.5).

As we should anticipate, the limiting stable random variable $\mathbf{S}^v(1)$ has the same tail-probability asymptotics as the service-time cdf G, i.e.,

$$P\big(\mathbf{S}^v(1) > t\big) \sim \gamma t^{-\alpha} \quad \text{as } t \to \infty, \tag{3.11}$$

as can be seen by combining (5.12) in section 4.5 of [20] with (3.5) and (3.6) here.

Assuming that the burstiness in the service times dominates the burstiness in the arrival process, it will be natural to have condition (3.2), as we have assumed. We want condition (3.2) in order to get a limit process without negative jumps. Having a limit process without negative jumps is not needed for theorem 2.1; we need condition (3.2) now in order to have a tractable steady-state distribution. We exploit that tractability further in the next theorem.

In the $GI/GI/1/K$ special case (which we do not require), if condition (2.9) holds, then the limit processes $\mathbf{S}^u$ and $\mathbf{S}^v$ are independent processes that must be either Brownian motions or stable Lévy motions, with condition (3.1) implying that $\mathbf{S}^v$ must be a stable Lévy motion. That is sufficient to have conditions (2.10) and (2.11) in theorem 2.1 hold. In this $GI/GI/1/K$ context, the assumed limit in (3.2) holds if the stable index of $\mathbf{S}^v$ is strictly smaller than the stable index of $\mathbf{S}^u$. In the $GI/GI/1/K$ context, condition (3.2) holds if the interarrival time cdf, say F, satisfies

$$t^{\alpha} F^{c}(t) \to 0 \quad \text{as } t \to \infty; \tag{3.12}$$

see [19, theorem 7]. In the $GI/GI/1/K$ case it suffices for the interarrival time to have finite variance. Given that $\mathbf{S}_n^u \Rightarrow 0\mathbf{e}$, conditions (2.9)–(2.11) in theorem 2.1 are satisfied with the limit processes $\mathbf{S}^v$ and $\mathbf{S}^u$ there being the stable Lévy motion and $0\mathbf{e}$, respectively.

A limit for the loss proportions thus follows from corollary 2.1, with the scaling by $c_n \equiv n^{1/\alpha}$ for $1 < \alpha < 2$.

As described in [20, section 8.5.2], when $\eta > 0$ (corresponding to $\rho_n < 1$), the content portion of the totally skewed ($\beta = 1$) limiting RSLM has a relatively tractable steady-state distribution. (That property extends to general Lévy processes without negative jumps.) In particular, $\mathbf{Q}(t) \Rightarrow \mathbf{Q}(\infty)$ as $t \to \infty$, where $\mathbf{Q}(\infty)$ has cdf H^{∞} given in (3.7)–(3.9). As shown in [20], it is not difficult to compute numerical values of the cdf H^{∞} by numerically inverting its Laplace transform. For more on the cdf H^{∞}, see [5, p. 191]. □

We now go beyond [20] to determine the steady-state loss proportion for the RSLM. we can exploit the fact that our totally-skewed stable Lévy motion can be represented as a deterministic negative drift plus a pure-jump process, with only positive jumps. The rate at which jumps of various sizes occur is determined by the *Lévy measure* of the stable Lévy motion; e.g., see [2], sections 8.2 and 8.3 of Bingham et al. [3], chapter XVII of Feller [8], chapters 3 and 7 of Gnedenko and Kolmogorov [9] and section 2.4 of the Internet supplement to [20]. In particular, jumps of at least size x occur according to a Poisson process at rate $\nu([x, \infty))$, where ν is the Lévy measure.

In our case, because of (3.11), the Lévy measure is

$$\nu\big([x, \infty)\big) = \frac{\gamma}{x^{\alpha}}; \tag{3.13}$$

see [2, p. 217] and [3, theorem 8.2.1]. It is significant that the two parameters α and γ in (3.13) are both determined by the initial tail-probability asymptotics for the service-time cdf G in (3.1).

We exploit the Lévy-process structure to establish the following result, proved in section 7.

Theorem 3.2. A limiting loss rate exists for the reflected stable Lévy motion obtained as a limit in theorem 3.1 and its value is

$$\pi \equiv \lim_{t\to\infty} \frac{\mathbf{L}(t)}{t} = \frac{\gamma\alpha}{(\alpha-1)} \int_0^{\kappa} (\kappa - x)^{1-\alpha} h(x)\,\mathrm{d}x, \tag{3.14}$$

for γ and α in (3.1), where h is the density of the cdf H of $\mathbf{Q}(\infty)$ determined in (3.7)–(3.9).

Combining (1.6), (2.18) and (3.14), we obtain the heavy-traffic approximation. Paralleling (2.28) for the Brownian case, in this heavy-tailed case the heavy-traffic approximation (1.6) becomes

$$\pi_n \approx \frac{\pi}{n^{(\alpha-1)/\alpha}} \tag{3.15}$$

for α in (3.1) and π in (3.14).

It now remains to compute or further approximate π. As indicated in [20, section 8.6.1], we can approximate the limiting RSLM by the queue-length process of an $M/GI/1/K$ queue, using a second heavy-traffic limit to determine the parameters of the $M/GI/1/K$ queue. We then can calculate the desired loss proportion in the $M/GI/1/K$ queue by numerical transform inversion from the Pollaczek–Khintchine transform (exploiting relations (4.1) and (4.2) in the next section).

Another attractive way to generate an approximation is to consider the asymptotic behavior as $\kappa \to \infty$. For that purpose, let π^{κ} denote the RSLM loss proportion as a function of the upper barrier κ.

Theorem 3.3. As $\kappa \to \infty$ for π^{κ} in (3.14),

$$\kappa^{\alpha-1}\pi^{\kappa} \to \frac{\gamma\alpha}{\alpha-1}, \tag{3.16}$$

independent of the drift η.

From theorem 3.3, we obtain the approximation

$$\pi^{\kappa} \approx \frac{\gamma\alpha}{(\alpha-1)\kappa^{\alpha-1}}. \tag{3.17}$$

Combining (3.15) and (3.17), we obtain

$$\pi_n \approx \frac{\gamma\alpha}{(\alpha-1)n^{(\alpha-1)/\alpha}\kappa^{\alpha-1}}. \tag{3.18}$$

Letting $\kappa = K_n/n^{1/\alpha}$ in (3.18), we obtain the further approximation

$$\pi_n \approx \frac{\gamma\alpha}{(\alpha-1)K_n^{\alpha-1}}. \tag{3.19}$$

Remark 3.1. Interchanging the order of the limits. As noted after corollary 13 of Whitt [19], in the infinite-capacity heavy-tailed case, the order of the limits for the tail probabilities can be interchanged: In the infinite-capacity case,

$$\begin{aligned} P\big(Q_n^{\infty}(nt) > K_n\big) &= P\big(n^{-1/\alpha}Q_n^{\infty}(nt) > n^{-1/\alpha}K_n\big) \\ &\to P\big(\mathbf{Q}^{\infty}(t) > \kappa\big) && \text{as } n \to \infty \\ &\to P\big(\mathbf{Q}^{\infty}(\infty) > \kappa\big) = H^{\infty,c}\bigg(\frac{\kappa}{\xi}\bigg) && \text{as } t \to \infty \\ &\sim \frac{\xi^{\alpha-1}}{\Gamma(2-\alpha)\kappa^{\alpha-1}} = \frac{\gamma}{\eta(\alpha-1)\kappa^{\alpha-1}} && \text{as } \kappa \to \infty. \end{aligned} \tag{3.20}$$

Those limits for the infinite-capacity system can be used as approximations for the loss proportion in the finite-capacity system. From our analysis, we see that the last expression in (3.20) differs from the finite-capacity limit in (3.16) only by the constant factor α/η. However, that difference is likely to be more significant than in the Brownian case because the rest of the loss proportion has a power decay instead of an exponential decay.

Our main point, though, is that the triple limit in (3.20), which is in the order $n \to \infty$, $t \to \infty$ and then $\kappa \to \infty$, agrees with the limit in the alternative order $t \to \infty$, $\kappa \to \infty$ and $n \to \infty$ (with $\rho_n \uparrow 1$ such that (2.7) holds) obtained from [6]. In contrast, in the finite-capacity case, theorem 3.3 and approximation (3.19) do *not* agree with the corresponding triple limit in the order $t \to \infty$, $x \to \infty$ and $n \to \infty$ (with $\rho_n \uparrow 1$ such that (2.7) and (2.8) hold) for the $GI/GI/1/K$ model established by Jelenković [12]. Our formulas (3.16) and (3.19) have an extra factor α in the numerator. Jelenković [12] actually considers a different model, which is model 1 in section 6 here, but that model has the same heavy-traffic limit, and so the same triple limit estasblished in theorems 3.1, 3.2 and 3.3.

The discrepancy raises the possibility that the expression for π in (3.14) is in error, having an extra factor α, but we have been unable to identify an error.

4. The $M/GI/1/K$ model

In this section we establish the local heavy-traffic limit for the $M/GI/1/K$ model, directly establishing a heavy-traffic limit for the long-run loss proportion. In doing so, we return to the light-tailed setting in which the service time has finite variance.

To establish our local heavy-traffic limit, we draw upon established theory of Takagi in [16, section 5.1]. We let K be the size of the waiting room, excluding the server. Thus our $K+1$ is Takagi's K.

Following Takagi [16], let the Poisson arrival rate be λ and let the service-time have cumulative distribution function B with finite mean b. Let the traffic intensity (offered load) be $\rho \equiv \lambda b$. Let P_k denote the steady-state probability that there are k customers in the system at an arbitrary time. By the Poisson-Arrivals-See-Time-Averages (PASTA) property, that is also the probability seen by an arrival. Thus the blocking probability, say P_B, is just P_{K+1}. Let π_k, $0 \leqslant k \leqslant K$, denote the steady-state probability of the queue length just after departures. (Necessarily, $\pi_{K+1} = 0$, whereas $P_{K+1} = P_B > 0$.)

We draw upon two important results. The first is the Cooper–Gebhardt relation between π_0 and P_B, namely,

$$P_B = 1 - \frac{1}{\pi_0 + \rho}; \tag{4.1}$$

see [16, (1.18b) on p. 202]. The second is the proportionality relation between the steady-state probabilities π_k and the associated steady-state probabilities in the infinite-waiting-room $M/GI/1/\infty$ model, denoted by π_k^∞,

$$\pi_k = \frac{\pi_k^\infty}{\sum_{j=0}^{K} \pi_j^\infty}, \quad 0 \leqslant k \leqslant K; \tag{4.2}$$

see [16, (1.23) on p. 205]. Clearly, (4.2) requires $\rho < 1$.

We apply these results and the known heavy-traffic limit for the $M/GI/\infty$ model to establish the desired local heavy-traffic limit for the $M/GI/1/K$ model. At this point we require that the service-time cdf B have a finite squared coefficient of variation (SCV, variance divided by the square of the mean) c_s^2.

To formulate the limit, we consider a sequence of $M/GI/1/K$ models indexed by n. We let the service-time distribution remain fixed and make the Poisson arrival rate change with n. We assume that the associated sequence of traffic intensities $\{\rho_n \colon n \geqslant 1\}$ and waiting-room sizes $\{K_n \colon n \geqslant 1\}$ increase as n increases, satisfying

$$\sqrt{n}(1 - \rho_n) \to \eta \quad \text{as } n \to \infty \tag{4.3}$$

and

$$\frac{K_n}{\sqrt{n}} \to \kappa \quad \text{as } n \to \infty. \tag{4.4}$$

Theorem 4.1. For the sequence of $M/GI/1/K$ models specified above, suppose that (4.3) and (4.4) hold with $0 < \eta < \infty$ and $0 < \kappa < \infty$, and suppose that the fixed service-time cdf B has a finite SCV c_s^2. Then

$$\sqrt{n} P_{b,n} \to \frac{\eta \mathrm{e}^{-2\eta\kappa/\sigma^2}}{1 - \mathrm{e}^{-2\eta\kappa/\sigma^2}}, \tag{4.5}$$

where

$$\sigma^2 = 1 + c_s^2. \tag{4.6}$$

Proof. By (4.1),

$$\sqrt{n} P_{b,n} = \frac{\sqrt{n}\pi_{n,0} - \sqrt{n}(1 - \rho_n)}{\pi_{n,0} + \rho_n}. \tag{4.7}$$

However, since $\pi^{\infty}_{n,0} = 1 - \rho_n$, by (4.2),

$$\sqrt{n}\pi_{n,0} = \frac{\sqrt{n}(1-\rho_n)}{\sum_{j=0}^{K_n} \pi^{\infty}_{n,j}}. \tag{4.8}$$

The proof is completed by applying (4.3), (4.4) and the known limit for the sequence of steady-state cdf's in the associated sequence of $M/GI/1/\infty$ models, the last to treat the denominator in (4.8). Specifically, we can apply a Taylor series expansion in the Polaczek–Khintchine characteristic function for the steady-state queue length, as was done on p. 168 of Gnedenko and Kovalenko [10], invoking the continuity theorem for characteristic functions, theorem 2 on p. 508 of Feller [8]. □

5. The dual $GI/M/1/K$ model

In this section we apply the result in the previous section for the $M/GI/1/K$ model with $\rho < 1$ to obtain a corresponding result for the "dual" $GI/M/1/K$ model with $\rho > 1$. We call it the dual model because we simply switch the role of the interarrival-time and service-time distributions.

Let ν denote the mean number of customers served in a busy period. We rely on the following connection between the $GI/M/1/K$ model and the associated "dual" $M/GI/1/K+1$ model.

Theorem 5.1. For the dual $GI/M/1/K$ and $M/GI/1/K$ models,

$$P_B(GI/M/1/K) = \frac{1}{\nu(M/GI/1/K+1)} = \pi_0(M/GI/1/K+1). \tag{5.1}$$

Proof. We use a regenerative argument to express the $GI/M/1/K$ blocking probability as the reciprocal of the expected number of arrivals between successive overflows. We then observe that the number of arrivals between successive overflows is distributed exactly the same as the number of customers served in a busy period in the dual $M/GI/1/K+1$ queue (with 1 more waiting space). Finally, we use another regenerative argument to equate the probability of emptiness just after a departure to the reciprocal of the expected number of customers served in a busy period. □

We now apply theorem 5.1 to obtain the local heavy-traffic limit for $GI/M/1/K$. We again consider a sequence of queueing models indexed by n, but now we fix the interarrival-time cdf, letting it be B with mean b and SCV c_a^2, and we let the service rate depend upon n. Now, instead of (4.3), we assume that

$$\sqrt{n}(\rho_n - 1) \to \eta \quad \text{as } n \to \infty \text{ for } 0 < \eta < \infty. \tag{5.2}$$

Theorem 5.2. For the sequence of $GI/M/1/K$ models specified above, suppose that (5.2) and (4.4) hold with $0 < \eta < \infty$ and $0 < \kappa < \infty$, and suppose that the fixed interarrival-time cdf B has a finite SCV c_a^2. Then

$$\sqrt{n}P_{b,n} \to \frac{\eta \mathrm{e}^{2\eta\kappa/\sigma^2}}{\mathrm{e}^{2\eta\kappa/\sigma^2} - 1}, \tag{5.3}$$

where

$$\sigma^2 = c_a^2 + 1. \tag{5.4}$$

Proof. Because of theorem 5.1, it suffices to establish the limit for the sequence of probabilities $\pi_{n,0}$ in the $M/GI/K + 1$ queues with traffic intensities $1/\rho_n$. However, since $\rho_n > 1$ for all n sufficiently large n, $1/\rho_n < 1$ for all sufficiently large n. Hence we can apply (4.2) for $k = 0$ and $K_n + 1$. As in the proof of theorem 4.1, we obtain

$$\sqrt{n}\pi_{n,0}(M/GI/1/K_n + 1) \to \frac{\eta}{1 - \mathrm{e}^{-2\eta\kappa/\sigma^2}}. \tag{5.5}$$

Multiplying through in (5.5) above and below by $\mathrm{e}^{2\eta\kappa/\sigma^2}$, we obtain the desired limit in (5.3). □

6. Related finite-capacity models

In this section we consider other finite-capacity single-server queueing models. Our main goal is to show that the same heavy-traffic stochastic-process limits hold for these related models, with the identical limit process. Unfortunately, however, we sometimes need to impose the extra condition that key limit processes almost surely have continuous sample paths. It remains to determine if these extra conditions can be removed.

We start by defining several truncation and loss (overflow) functions. For $y \in \mathbb{R}$, let

$$[y]_a \equiv \max\{y, a\}, \qquad [y]^b \equiv \min\{y, b\},$$
$$[y]_a^b \equiv \big[[y]_a\big]^b, \qquad \lambda^b(y) \equiv y - [y]^b \quad \text{and} \quad \lambda_a(y) \equiv [y]_a - y.$$

Our first alternative model is the $G/G/1/K$ queue with *uniformly bounded actual waiting time*, as in [7, chapter III.4]. For all models, we start with the same basic sequence $\{(U_k, V_k)\}$ as in section 2. Let W_k^1 be the waiting time (before beginning service) of the kth customer; let $Y_k^{u,1}$ be the lost service time (if any) associated with the kth customer and let $Y_k^{l,1}$ be the idle time associated with the kth customer. These processes can be defined recursively by

$$\begin{aligned} W_{k+1}^1 &\equiv \big[W_k^1 + V_k - U_{k+1}\big]_0^K, && k \geqslant 2, \\ Y_k^{u,1} &\equiv \lambda^K\big(W_k^1 + V_k - U_{k+1}\big), && k \geqslant 1, \\ Y_k^{l,1} &\equiv \lambda_0\big(W_k^1 + V_k - U_{k+1}\big), && k \geqslant 1, \end{aligned} \tag{6.1}$$

with $W_1^1 \equiv 0$, assuming that the first customer arrives at time U_1 to find an empty system. Other interpretations of the same model are obtained by changing the interpretation of the variables U_k and V_k; e.g., V_k may be regarded as a one-period input, while U_k may be regarded as a potential one-period output. This model is clearly quite tractable, because the two-sided reflection map can be applied directly. This model was discussed in [1,19] and [20, section 2.3 and chapters 5 and 8]. (However, the results in section 3 here are new for this model.) This is the model considered by Jelenković [12] too. It is discussed again here for comparison purposes.

Our second model is the *finite dam* or, equivalently, the discrete-time embedded process in the $G/G/1/K$ queue with *uniformly bounded virtual waiting time*, as in [7, chapter III.5]. Instead of (6.1), the definitions now are

$$\begin{aligned} W_{k+1}^2 &\equiv \big[\big[W_k^2 + V_k\big]^K - U_{k+1}\big]_0, & k &\geqslant 2,\\ Y_k^{u,2} &\equiv \lambda^K\big(W_k^2 + V_k\big), & k &\geqslant 1,\\ Y_k^{l,2} &\equiv \lambda_0\big(W_k^2 + V_k - U_{k+1} - \lambda^K\big(W_k^2 + V_k\big)\big), & k &\geqslant 1, \end{aligned} \tag{6.2}$$

with $W_1^1 \equiv 0$. The finite dam has extra interest, because Zwart [23] has connected it to a general finite-capacity fluid model.

A dual to the finite dam has the decrease each period occur before the increase, leading to

$$\begin{aligned} W_{k+1}^3 &\equiv \big[\big[W_k^3 - U_{k+1}\big]_0 + V_k\big]^K, & k &\geqslant 2,\\ Y_k^{u,3} &\equiv \lambda^K\big(\big[W_k^3 - U_{k+1}\big]_0 + V_k\big), & k &\geqslant 1,\\ Y_k^{l,3} &\equiv \lambda_0\big(W_k^3 - U_{k+1}\big), & k &\geqslant 1, \end{aligned} \tag{6.3}$$

with $W_1^1 \equiv 0$. Model 3 can be obtained from model 2 by looking at the empty space $K - W_k^2$ and switching the roles of the interarrival times and service times.

Closely related to model 3 is the $G/G/1/K$ model with *bounded actual sojourn time*. In this model, W_k^4 represents the (possibly truncated) sojourn time (waiting time plus service time) of customer k. Then

$$\begin{aligned} W_{k+1}^4 &\equiv \big[\big[W_k^4 - U_{k+1}\big]_0 + V_{k+1}\big]^K, & k &\geqslant 2,\\ Y_k^{u,4} &\equiv \lambda^K\big(\big[W_k^4 - U_{k+1}\big]_0 + V_{k+1}\big), & k &\geqslant 1,\\ Y_k^{l,4} &\equiv \lambda_0\big(W_k^4 - U_{k+1}\big), & k &\geqslant 1, \end{aligned} \tag{6.4}$$

with $W_1^4 \equiv V_1$. Model 4 differs from model 3 only by the initial condition and the shifted index in the service times.

Let $U^{\uparrow} \equiv \max\{U_1, \dots, U_k\}$ and similarly for V_k. Let the cumulative overflow processes be defined by

$$\begin{aligned} L_k^{u,i} &\equiv Y_1^{u,i} + \cdots + Y_k^{u,i},\\ L_k^{l,i} &\equiv Y_1^{l,i} + \cdots + Y_k^{l,i}, \quad k \geqslant 1. \end{aligned} \tag{6.5}$$

It is elementary that we can write, for each i,

$$W_{k+1}^i - W_k^i = V_k - U_{k+1} - Y_k^{u,i} + Y_k^{l,i}, \quad k \geqslant 1, \tag{6.6}$$

so that

$$W_k^i = S_{k-1}^v - S_k^u - L_{k-1}^{u,i} + L_{k-1}^{l,i}, \quad k \geqslant 1. \tag{6.7}$$

As a consequence of (6.7),

$$\begin{aligned} W_k^1 - W_k^2 &= \big(L_{k-1}^{u,2} - L_{k-1}^{u,1}\big) - \big(L_{k-1}^{l,2} - L_{k-1}^{l,1}\big), \\ W_k^3 - W_k^1 &= \big(L_{k-1}^{l,3} - L_{k-1}^{l,1}\big) - \big(L_{k-1}^{u,3} - L_{k-1}^{u,1}\big). \end{aligned} \tag{6.8}$$

By induction from the definitions, we obtain the following basic comparison lemma, which is proved in section 7.

Lemma 6.1. For all k,

$$\begin{aligned} &W_k^2 \leqslant W_k^1 \leqslant W_k^2 + V_{k-1}^{\uparrow}, \\ &W_k^3 \geqslant W_k^1 \geqslant W_k^3 - U_k^{\uparrow}, \\ &L_k^{u,2} \geqslant L_k^{u,1} \quad \text{and} \quad L_k^{l,2} \geqslant L_k^{l,1}, \\ &L_k^{u,3} \geqslant L_k^{u,1} \quad \text{and} \quad L_k^{l,3} \geqslant L_k^{l,1}. \end{aligned} \tag{6.9}$$

Remark 6.1. Two-sided bounds for the overflow processes. Unlike for the waiting times, it is not possible to obtain good general two-sided bounds for the overflow processes. In general, the cumulative-overflow processes for models 2 and 3 can be much larger than the overflow processes for model 1. For example, we can obtain the bound

$$L_k^{u,1} \leqslant L_k^{u,2} \leqslant L_k^{u,1} + S_k^v, \tag{6.10}$$

but trivially $L_k^{u,i} \leqslant S_k^v$ for all i. We can obtain better bounds by exploiting the number of times that the waiting-time sequences successively hit the upper and lower barriers; see the proof of theorem 6.1 below.

Now, just as in section 2, consider sequences of these queueing models indexed by n. We again use definitions (2.1)–(2.4). For each model i, let scaled random functions be defined by

$$\begin{aligned} \mathbf{W}_n^i(t) &\equiv \frac{W_{n,\lfloor nt\rfloor}^i}{c_n}, \\ \mathbf{L}_n^{u,i}(t) &\equiv \frac{L_{n,\lfloor nt\rfloor}^{u,i}}{c_n}, \\ \mathbf{L}_n^{l,i}(t) &\equiv \frac{L_{n,\lfloor nt\rfloor}^{l,i}}{c_n}, \quad t \geqslant 0. \end{aligned} \tag{6.11}$$

We now state analogs of theorem 2.1 for these related finite-capacity models. Limits hold for $\mathbf{L}_n^{l,i}$ just like for $\mathbf{L}_n^{u,i}$, but we only state results for $\mathbf{L}_n^{u,i}$. For that purpose, let C denote the subset of continuous functions in D.

Theorem 6.1. Suppose that the conditions of theorem 2.1 hold, again with the systems starting out empty. Then

(a)

$$\big(\mathbf{W}_n^1, \mathbf{L}_n^{u,1}\big) \Rightarrow (\mathbf{Q}, \mathbf{L}) \quad \text{in } D^2, \tag{6.12}$$

where $(\mathbf{Q}, \mathbf{L})$ is the limit process in (2.12).

(b) If $P(\mathbf{S}^v \in C) = 1$, then

$$\big(\mathbf{W}_n^1, \mathbf{W}_n^2, \mathbf{L}_n^{u,1}, \mathbf{L}_n^{u,2}\big) \Rightarrow (\mathbf{Q}, \mathbf{Q}, \mathbf{L}, \mathbf{L}) \quad \text{in } D^4. \tag{6.13}$$

(c) If $P(\mathbf{S}^u \in C) = 1$, then

$$\big(\mathbf{W}_n^1, \mathbf{W}_n^3, \mathbf{W}_n^4, \mathbf{L}_n^{u,1}, \mathbf{L}_n^{u,3}, \mathbf{L}_n^{u,4}\big) \Rightarrow (\mathbf{Q}, \mathbf{Q}, \mathbf{Q}, \mathbf{L}, \mathbf{L}, \mathbf{L}) \quad \text{in } D^6. \tag{6.14}$$

(d) If $P(\mathbf{S}^u \in C) = P(\mathbf{S}^v \in C) = 1$, then

$$\big(\mathbf{W}_n^1, \mathbf{W}_n^2, \mathbf{W}_n^3, \mathbf{W}_n^4, \mathbf{L}_n^{u,1}, \mathbf{L}_n^{u,2}, \mathbf{L}_n^{u,3}, \mathbf{L}_n^{u,4}\big) \Rightarrow (\mathbf{Q}, \mathbf{Q}, \mathbf{Q}, \mathbf{Q}, \mathbf{L}, \mathbf{L}, \mathbf{L}, \mathbf{L}) \quad \text{in } D^8. \tag{6.15}$$

Under the assumptions of theorem 6.1, we have analogs of corollary 2.1 for the four models in (6.1)–(6.4). Since the argument is essentially the same, we only state the main part of the result, omitting the proof. Let $C_n(t)$ denote the cumulative input in the time interval $[0, t]$ with index n; i.e.,

$$C_n(t) \equiv \sum_{j=1}^{A_n(t)} V_{n,j}, \quad t \geqslant 0. \tag{6.16}$$

Then the proportion of input lost over the interval $[0, t]$ in model i with index n is

$$\Pi_n^i(t) \equiv \frac{L_{n,A_n(t)}^{u,i}}{C_n(t)}, \quad t \geqslant 0, \tag{6.17}$$

provided that $C_n(t) > 0$, with $\Pi_n^i(t) \equiv 0$ if $C_n(t) = 0$. Paralleling (2.14), let the associated scaled processes be defined by

$$\mathbf{\Pi}_n^i(t) \equiv \frac{n\Pi_n^i(nt)}{c_n}, \quad t \geqslant 0. \tag{6.18}$$

Corollary 6.1. Under the assumptions of theorem 6.1 (which depend on model i),

$$\mathbf{\Pi}_n^i \Rightarrow \mathbf{\Pi} \quad \text{in } D\big((0, \infty), \mathbb{R}\big), \tag{6.19}$$

where $\mathbf{\Pi}(t) \equiv \mathbf{L}(t)/t$, $t \geqslant 0$.

As described in [23], it is not difficult to give expressions for the steady-state loss in the $GI/GI/1/K$ special case by exploiting regenerative arguments. So now consider a single $GI/GI/1/K$ model. Let U and V be a generic interarrival time and service time with $EV = 1$. Let W_∞^i and $Y_\infty^{u,i}$ be the steady-state quantities; see [7] for discussion of existence. Since $EV = 1$, $EY_\infty^{u,i}$ is the steady-state loss proportion, the analog of the blocking probability π in section 2.

For any nonnegative random variable V, let V_e be a random variable with the associated *stationary-excess* (or residual lifetime) distribution, i.e.,

$$P(V_e \leqslant x) \equiv \frac{1}{EV}\int_0^x P(V > y)\,\mathrm{d}y. \tag{6.20}$$

We will use the fact that $EV_e = E[V^2]/2EV$. (For the service time, $EV = 1$.)

By an elementary regenerative argument, we obtain the following expressions for the steady-state loss proportions.

$$\begin{aligned} EY_\infty^{u,1} &= \int_0^\infty P\big(W_\infty^1 + V - U > K + x\big)\,\mathrm{d}x = P\big(W_\infty^1 + V_e - U > K\big), \\ EY_\infty^{u,2} &= \int_0^\infty P\big(W_\infty^2 + V > K + x\big)\,\mathrm{d}x = P\big(W_\infty^2 + V_e > K\big), \\ EY_\infty^{u,3} = EY_\infty^{u,4} &= \int_0^\infty P\big(\big[W_\infty^3 - U\big]_0 + V > K + x\big)\,\mathrm{d}x = P\big(\big[W_\infty^3 - U\big]_0 + V_e > K\big) \end{aligned} \tag{6.21}$$

where W_∞^i, U, V and V_e are mutually independent in the expressions on the right. The independence structure makes it possible to exploit (6.21) for computations, but that is not our purpose here.

Zwart [23] has shown that, just as in section 4, the finite-dam formulas simplify greatly in the $M/GI/1/K$ special case when U has an exponential distribution. Let W_∞^∞ be the steady-state waiting time in the $M/GI/1/\infty$ model. In particular, from (3.6) and (3.9) of [23], we have

$$P\big(W_\infty^2 \leqslant x\big) = \frac{P(W_\infty^\infty \leqslant x)}{P(W_\infty^\infty \leqslant K)} \tag{6.22}$$

and

$$EY_\infty^{u,2} = \frac{(1-\rho)P(W_\infty^\infty > K)}{\rho P(W_\infty^\infty \leqslant K)} \tag{6.23}$$

when $\rho < 1$. Thus, for the $M/GI/1/K$ finite dam, we can obtain analogs of section 4, but we do not state these results. Moreover, we can apply [5] to obtain heavy-traffic limits in the case of heavy-tailed service-time distributions. For the case $\rho = 1$, the limits are given in proposition 7.1 of Zwart [23].

We can also apply the expressions in (6.21) to obtain heuristic heavy-traffic approximations for the steady-state loss proportions. For that purpose, we again consider a sequence of queueing models indexed by n, where (2.7) and (2.8) hold. Consider the case of the finite dam. For each n, the model is a $GI/GI/1/K$ finite dam. Let the service-time distribution be fixed with mean 1 and finite SCV c_s^2; that implies that $EV_e = (c_s^2 + 1)/2$. Consistent with theorem 6.1, but not directly implied by it, we now assume that $c_n^{-1} W_{n,\infty}^2 \Rightarrow \mathbf{Q}(\infty)$, where $\mathbf{Q}(\infty)$ has the steady-state distribution of the limit process $\mathbf{Q}$. We now assume that the distribution of $\mathbf{Q}(\infty)$ is absolutely continuous with density $f_{\mathbf{Q}(\infty)}$, as is the case with RBM.

By (6.21),

$$\begin{aligned} c_n E\bigl[Y_{n,\infty}^{u,2}\bigr] &= c_n P\bigl(W_{n,\infty}^2 + V_e > K_n\bigr) \\ &= c_n P\bigl(c_n^{-1} W_{n,\infty}^2 > c_n^{-1} K_n - c_n^{-1} V_e\bigr) \approx c_n P\bigl(\mathbf{Q}(\infty) > \kappa - c_n^{-1} V_e\bigr) \\ &\approx c_n f_{\mathbf{Q}(\infty)}(\kappa) c_n^{-1} E[V_e] = f_{\mathbf{Q}(\infty)}(\kappa) \frac{c_s^2 + 1}{2}. \end{aligned} \tag{6.24}$$

Note, however, that the scaling by c_n in (6.24) is *inconsistent* with the scaling by n/c_n in corollary 2.1 unless $c_n/\sqrt{n} \to c$ as $n \to \infty$ for $0 < c < \infty$. Of course, $c_n = \sqrt{n}$ is a common case, but it is interesting to note the inconsistency in other cases. Since heuristics are used at this point, there is no logical error. There are some candidate explanations for the inconsistency: When $c_n/\sqrt{n} \to c$ for $0 < c < \infty$ fails to hold, we are likely to have (but do not necessarily have) $E[V^2] = \infty$. Also, when $c_n/\sqrt{n} \to c$ fails to hold, the distribution of $\mathbf{Q}(\infty)$ is likely not to be absolutely continuous.

In the standard Brownian case, $c_n = \sqrt{n}$ and $\mathbf{Q}$ is RBM. For the $M/GI/1$ finite dam, approximation (6.24) is consistent with (6.23), (2.27) and (2.28). It is also consistent with proposition 7.1 of Zwart [23]. However, (6.24) is inconsistent with (2.27) and (2.28) for the $GI/GI/1/K$ model with $c_a^2 \neq 1$ because $\sigma^2 = c_a^2 + c_s^2$.

Corresponding heuristics follow from the other formulas in (6.21), but it is not clear how to treat the variable U appearing there. For example, a corresponding approximation for $EY_\infty^{u,1}$ is

$$\begin{aligned} c_n E\bigl[L_{n,\infty}^1\bigr] &= c_n P\bigl(W_{n,\infty}^1 + V_e > K_n + \rho_n^{-1} U\bigr) \\ &= c_n P\bigl(c_n^{-1} W_{n,\infty}^1 > c_n^{-1} K_n - c_n^{-1}\bigl[V_e - \rho_n^{-1} U\bigr]_0\bigr) \\ &\approx c_n P\bigl(\mathbf{Q}(\infty) > \kappa - c_n^{-1}\bigl[V_e - \rho_n^{-1} U\bigr]_0\bigr) \\ &\approx c_n f_{\mathbf{Q}(\infty)}(\kappa) c_n^{-1} E\bigl([V_e - U]_0\bigr) = f_{\mathbf{Q}(\infty)}(\kappa) E\bigl([V_e - U]_0\bigr), \end{aligned} \tag{6.25}$$

but the term $E([V_e - U]_0)$ is inconsistent with the limit for $\mathbf{L}_n^1$ in (6.12) even when $c_n = \sqrt{n}$ and $\mathbf{Q}(\infty)$ is RBM, since it does not depend on the interarrival-time and service-time distributions only through their first two moments.

7. Proofs

In this section we present the omitted proofs.

Proof of theorem 2.1. The proof is similar to the proof of theorem 9.3.4 of [20] for the infinite-capacity case. (Here the drift is $-\eta_n$ instead of $+\eta_n$.) First, as in [20], the conditions imply that

$$(\mathbf{A}_n, \mathbf{N}_n) \Rightarrow \left(-\mathbf{S}^u, -\mathbf{S}^v\right) \quad \text{in } D^2. \tag{7.1}$$

Then, just as in (3.23)–(3.27) on pp. 298–299 of [20], we deduce that

$$\mathbf{Q}_n = \phi_{0,K_n/c_n}\left(\mathbf{A}_n - \mathbf{N}_n \circ \widehat{\mathbf{B}}_n - \eta_n \mathbf{e}\right), \tag{7.2}$$

and

$$\mathbf{L}_n = \psi^U_{K_n/c_n}\left(\mathbf{A}_n - \mathbf{N}_n \circ \widehat{\mathbf{B}}_n - \eta_n \mathbf{e}\right), \tag{7.3}$$

where $\circ$ is the composition map, $\phi_{0,\kappa}$ is the content portion of the two-sided reflection map, ψ^U_κ is the upper-boundary regulator portion of the two-sided reflection map and $\widehat{\mathbf{B}}_n$ is a scaled version of the cumulative busy time in the interval $[0, t]$, $B_n(t)$, in particular,

$$\widehat{\mathbf{B}}_n(t) \equiv \frac{B_n(nt)}{n}, \quad t \geqslant 0. \tag{7.4}$$

(Note that η_n in (2.7) is defined slightly differently from η_n in (3.11) [20, p. 295].) In that argument, we start with

$$\begin{aligned} Q_n(t) &= A_n(t) - N_n\big(B_n(t)\big) - L_n(t) \\ &= A_n(t) - N_n\big(B_n(t)\big) - L_n(t) + [e - B_n](t) - [e - B_n](t) \\ &= \phi_{0,K_n}\big(A_n - N_n \circ B_n - [e - B_n]\big)(t), \quad t \geqslant 0, \end{aligned} \tag{7.5}$$

and

$$L_n(t) = \psi^U_{K_n}\big(A_n - N_n \circ B_n - [e - B_n]\big)(t), \quad t \geqslant 0. \tag{7.6}$$

Then, as in (3.27) of [20, p. 299], we observe that

$$\left(\mathbf{A}_n - \mathbf{N}_n \circ \widehat{\mathbf{B}}_n - \eta_n \mathbf{e}\right)(t) = c_n^{-1}\big(A_n - N_n \circ B_n - [e - B_n]\big)(nt), \quad t \geqslant 0. \tag{7.7}$$

The reason that we can apply the two-sided reflection without doing any special modification at the upper barrier is that the arrival process is exogenous, producing autonomous arrivals, even though we do not have autonomous service; see the discussion in [20, section 10.1.2].

In the proof of theorem 9.3.4 of [20], we could first treat the virtual-waiting-time or workload process and, as a consequence, directly deduce that $\widehat{\mathbf{B}}_n \Rightarrow \mathbf{e}$. With the upper boundary, it is no longer easy to first treat the workload process, so we cannot directly deduce that $\widehat{\mathbf{B}}_n \Rightarrow \mathbf{e}$. However, we can use an alternative argument, as in the proof of 14.7.4 of [20]. We observe that the sequence $\{\widehat{\mathbf{B}}_n\}$ is necessarily tight because it is uniformly Lipschitz: For $0 < t_1 < t_2$,

$$\left|\widehat{\mathbf{B}}_n(t_2) - \widehat{\mathbf{B}}_n(t_1)\right| \leqslant |t_2 - t_1|; \tag{7.8}$$

see [20, theorem 11.6.3]. By Prohorov's theorem, theorem 11.6.1 of [20], tightness implies relative compactness. Thus, any subsequence of $\{\widehat{\mathbf{B}}_n\colon\ n \geqslant 1\}$ has a convergent subsequence.

We then consider a convergent subsequence: Suppose that $\widehat{\mathbf{B}}_{n_k} \Rightarrow \widehat{\mathbf{B}}$ as $n_k \to \infty$. Unlike the proof of theorem 14.7.4 of [20], we cannot proceed by establishing a functional weak law of large numbers (FWLLN) with scaling by n instead of by c_n, because the upper barriers have been scaled by c_n. However, we can establish a FCLT along the subsequence with scaling by c_n. In our single-server-queue setting, the key is to relate the cumulative busy time to the cumulative idle time: Let $I_n(t)$ be the cumulative server idle time in the interval $[0, t]$, i.e., $I_n(t) \equiv t - B_n(t)$, $t \geqslant 0$, and let

$$\mathbf{I}_n(t) \equiv \frac{I_n(nt)}{c_n}, \quad t \geqslant 0. \tag{7.9}$$

Since $I_n = e - B_n$, as in (7.5) and (7.6), we can represent $\mathbf{I}_n$ as

$$I_n(t) = \psi_0^L\big(A_n - N_n \circ B_n - [e - B_n]\big)(t), \quad t \geqslant 0. \tag{7.10}$$

Because of (7.7),

$$\mathbf{I}_n = \psi_0^L\big(\mathbf{A}_n - \mathbf{N}_n \circ \widehat{\mathbf{B}}_n - \eta_n \mathbf{e}\big). \tag{7.11}$$

Along the subsequence n_k, we have

$$\big(\mathbf{A}_{n_k}, \mathbf{N}_{n_k}, \widehat{\mathbf{B}}_{n_k}\big) \Rightarrow \big(-\mathbf{S}^u, -\mathbf{S}^v, \widehat{\mathbf{B}}\big). \tag{7.12}$$

Thus, by (7.11), (7.12) and the continuous-mapping theorem with the two-sided reflection map, we have

$$\mathbf{I}_{n_k} \Rightarrow \psi_0^L\big(\mathbf{S}^v \circ \widehat{\mathbf{B}} - \mathbf{S}^u - \eta \mathbf{e}\big). \tag{7.13}$$

Since $n/c_n \to \infty$, we thus have

$$\hat{\mathbf{I}}_{n_k} \to 0\mathbf{e}, \tag{7.14}$$

where $\hat{\mathbf{I}}_n(t) \equiv I_n(nt)/n$, $t \geqslant 0$. Since $B_n(t) = t - I_n(t)$, we conclude that $\widehat{\mathbf{B}} = \mathbf{e}$. Since that same limit is obtained for all subsequences, the entire sequence must converge to that limit; i.e., we must have $\widehat{\mathbf{B}}_n \Rightarrow \mathbf{e}$.

Given that $\widehat{\mathbf{B}}_n \Rightarrow \mathbf{e}$, we can apply the continuous-mapping theorem with the two-sided reflection map in the setting of (7.2) and (7.3) in order to obtain the desired result. We actually have a sequence of reflection maps, but it is not difficult to see that

$$\big(\phi_{0,\kappa_n}(x_n), \psi_0^L(x_n), \psi_{\kappa_n}^U(x_n)\big) \to \big(\phi_{0,\kappa}(x), \psi_0^L(x), \psi_\kappa^U(x)\big) \quad \text{in } D^3$$

whenever $x_n \to x$ in D and $\kappa_n \to \kappa > 0$ in $\mathbb{R}$. For that step, we can exploit the Lipschitz property

$$d_{M_1}\big(\phi_{0,\kappa_n}(x_n), \phi_{0,\kappa_n}(x)\big) \leqslant 2 d_{M_1}(x_n, x) \tag{7.15}$$

uniformly in κ_n; see [20, theorem 14.8.5]. Hence the proof is complete. □

Proof of corollary 2.1. By (7.1), $\hat{\mathbf{A}}_n \to \mathbf{e}$ as $n \to \infty$, where $\hat{\mathbf{A}}_n(t) \equiv [A_n(nt)]_1/n$, $t \geqslant 0$. By [20, theorem 11.4.5],

$$\left(\mathbf{L}_n, \hat{\mathbf{A}}_n\right) \Rightarrow (\mathbf{L}, \mathbf{e}) \quad \text{in } D^2. \tag{7.16}$$

By the continuous-mapping theorem with the function $h : D^2 \to D((0, \infty), \mathbb{R})$ defined by

$$h(x, y)(t) \equiv \frac{x(t)}{y(t)}, \quad t > 0, \tag{7.17}$$

drawing on theorem 13.3.2 of [20], we have the main limit in (2.15). The limit in (2.16) follows from the continuous-mapping theorem with the projection map. The limit in (2.17) follows directly from (2.16) and condition (2.7). The iterated double limit in (2.18) follows directly from (2.16) under the extra assumption. □

Proof of theorem 3.2. By (3.11), jumps of size at least x occur according to a Poisson process at rate $\nu([x, \infty)) = \gamma/x^{\alpha}$. Let J_x be a random variable having the distribution of the conditional size of a jump given that there is a jump of at least size x. Then

$$E[J_x] = \frac{\int_x^\infty y\nu(\mathrm{d}y)}{\nu([x, \infty))} = \frac{\alpha x}{\alpha - 1}. \tag{7.18}$$

Finally, we are ready to compute the steady-state loss proportion π for the RSLM. A regenerative argument shows that $\mathbf{L}(t)/t \Rightarrow \pi$ as $t \to \infty$. To calculate π we consider the system in steady state. Then

$$\begin{aligned} \pi &= \int_0^\kappa \nu(\kappa - x) E[J_{\kappa - x}]\, h(x)\, \mathrm{d}x \\ &= \int_0^\kappa \frac{\gamma\alpha}{(\alpha - 1)(\kappa - x)^{\alpha - 1}} h(x)\, \mathrm{d}x, \end{aligned} \tag{7.19}$$

where h is the density of the cdf H of $\mathbf{Q}(\infty)$ determined in (3.7)–(3.9). □

Proof of theorem 3.3. Note that

$$\kappa^{\alpha - 1}\pi^{\kappa} = \frac{\gamma\alpha}{\alpha - 1} \int_0^\kappa \left(1 - \frac{x}{\kappa}\right)^{1-\alpha} \left(\frac{h^\infty(x)}{H^\infty(K)}\right) \mathrm{d}x. \tag{7.20}$$

We break the integral into two parts and apply the dominated convergence theorem to each part. First, since $(1 - (x/\kappa))^{1-\alpha} \leqslant \epsilon^{1-\alpha}$ for $x \leqslant (1 - \epsilon)\kappa$, the integrand of

$$\int_0^\infty \mathbb{1}_{[0,(1-\epsilon)\kappa]}(x) \left(1 - \frac{x}{\kappa}\right)^{1-\alpha} h^\infty(x)\, \mathrm{d}x \tag{7.21}$$

is dominated by $\epsilon^{1-\alpha} h^\infty(x)$, which integrates to $\epsilon^{1-\alpha}$. Since the integrand of (7.21) converges to $h^\infty(x)$ pointwise as $\kappa \to \infty$, the integral in (7.21) converges to 1.

Now consider the other piece and make the change of variables $y = x/\kappa$ to get

$$\int_{(1-\epsilon)\kappa}^{\kappa} \left(1 - \frac{x}{\kappa}\right)^{1-\alpha} h^{\infty}(x)\,\mathrm{d}x = \int_{(1-\epsilon)}^{1} (1-y)^{1-\alpha} h^{\infty}(\kappa y)\kappa\,\mathrm{d}y. \tag{7.22}$$

However, by theorem 8.5.3 of [20], $h^{\infty}(\kappa y) \sim c(\kappa y)^{-\alpha}$ as $\kappa \to \infty$ for some positive constant c. Hence

$$h^{\infty}(\kappa y)\kappa \leqslant 2c\kappa(\kappa y)^{-\alpha} \quad \text{for all } \kappa \text{ sufficiently large.}$$

Hence we can apply the dominated convergence theorem to deduce that the integral in (7.22) converges to 0. □

Proof of lemma 6.1. We apply mathematical induction, starting with the waiting times. First, observe that $W_1^1 = W_1^2 = W_1^3 = 0$. We now show that the inequalities for the waiting times hold for index $k+1$ given that they hold for index k. For the first line,

$$\begin{aligned} W_{k+1}^2 &\equiv \left[\left[W_k^2 + V_k\right]^K - U_{k+1}\right]_0 \\ &\leqslant \left[\left[W_k^1 + V_k\right]^K - U_{k+1}\right]_0 \\ &\leqslant \left[W_k^1 + V_k - U_{k+1}\right]_0^K \equiv W_{k+1}^1 \end{aligned} \tag{7.23}$$

and

$$\begin{aligned} W_{k+1}^2 &\equiv \left[\left[W_k^2 + V_k\right]^K - U_{k+1}\right]_0 \\ &\geqslant \left[\left[W_k^1 - V_{k-1}^{\uparrow} + V_k\right]^K - U_{k+1}\right]_0 \\ &\geqslant \left[W_k^1 - V_{k-1}^{\uparrow} + V_k - U_{k+1}\right]_0^K - \left[V_k - V_{k-1}^{\uparrow}\right]_0 \\ &\geqslant \left[W_k^1 + V_k - U_{k+1}\right]_0^K - V_{k-1}^{\uparrow} - \left[V_k - V_{k-1}^{\uparrow}\right]_0 = W_{k+1}^1 - V_k^{\uparrow}. \end{aligned} \tag{7.24}$$

For the second line,

$$\begin{aligned} W_{k+1}^3 &\equiv \left[\left[W_k^3 - U_{k+1}\right]_0 + V_k\right]^K \\ &\geqslant \left[\left[W_k^1 - U_{k+1}\right]_0 + V_k\right]^K \\ &\geqslant \left[W_k^1 + V_k - U_{k+1}\right]_0^K \equiv W_{k+1}^1 \end{aligned} \tag{7.25}$$

and

$$\begin{aligned} W_{k+1}^3 &\equiv \left[\left[W_k^3 - U_{k+1}\right]_0 + V_k\right]^K \\ &\leqslant \left[\left[W_k^1 + U_k^{\uparrow} - U_{k+1}\right]_0 + V_k\right]^K \\ &\leqslant \left[\left[W_k^1 + U_k^{\uparrow} - U_{k+1} + V_k\right]_0^K + [U_{k+1} - U_k^{\uparrow}\right]_0 \\ &\leqslant \left[W_k^1 + V_k - U_{k+1}\right]_0^K + U_k^{\uparrow} + \left[U_{k+1} - U_k^{\uparrow}\right]_0 = W_{k+1}^1 + U_{k+1}^{\uparrow}. \end{aligned} \tag{7.26}$$

We now turn to the overflow processes. To start, note that

$$\begin{aligned}
L_1^{l,3} &= Y_1^{l,3} = \lambda_0(-U_2) = U_2 \geqslant \lambda_0(V_1 - U_2) = Y_1^{l,1} = L_1^{l,1},\\
L_1^{l,2} &= Y_1^{l,2} = \lambda_0\bigl(V_1 - U_2 - \lambda^K(V_1)\bigr) \geqslant \lambda_0(V_1 - U_2) = Y_1^{l,1} = L_1^{l,1},\\
L_1^{u,1} &= Y_1^{u,1} = \lambda^K(V_1 - U_2) \leqslant \lambda^K\bigl(V_1 - U_2 - Y_1^{l,2}\bigr) = Y_1^{u,3} = L_1^{u,3},\\
L_1^{u,1} &= Y_1^{u,1} = \lambda^K(V_1 - U_2) \leqslant \lambda^K(V_1) = Y_1^{u,2} = L_1^{u,2},
\end{aligned} \tag{7.27}$$

so that $L_1^{l,3} \geqslant L_1^{l,1}$, $L_1^{l,2} \geqslant L_1^{l,1}$, $L_1^{u,3} \geqslant L_1^{u,1}$ and $L_1^{u,2} \geqslant L_1^{u,1}$. To go from $k-1$ to k, observe that

$$L_k^{l,2} - L_k^{l,1} = L_{k-1}^{l,2} - L_{k-1}^{l,1} + Y_k^{l,2} - Y_k^{l,1} \geqslant 0 \tag{7.28}$$

because

$$Y_k^{l,2} = \lambda_0\bigl(W_k^2 + V_k - U_{k+1} - Y_k^{u,2}\bigr) \geqslant \lambda_0\bigl(W_k^1 + V_k - U_{k+1}\bigr) = Y_k^{l,1}. \tag{7.29}$$

Combining (6.8) and (7.28), we obtain

$$L_k^{u,2} - L_k^{u,1} = W_k^1 - W_k^2 + L_k^{l,2} - L_k^{l,1} \geqslant W_k^1 - W_k^2 \geqslant 0. \tag{7.30}$$

Next,

$$L_k^{u,3} - L_k^{u,1} = L_{k-1}^{u,3} - L_{k-1}^{u,1} + Y_k^{u,3} - Y_k^{u,1} \geqslant 0 \tag{7.31}$$

because

$$Y_k^{u,3} = \lambda^K\bigl(W_k^3 + V_k - U_{k+1} + Y_k^{l,3}\bigr) \geqslant \lambda^K\bigl(W_k^1 + V_k - U_{k+1}\bigr) = Y_k^{u,1}. \tag{7.32}$$

Combining (6.8) and (7.31), we obtain

$$L_k^{l,3} - L_k^{l,1} = W_k^3 - W_k^1 + L_k^{u,3} - L_k^{u,1} \geqslant W_k^3 - W_k^1 \geqslant 0. \tag{7.33}$$

□

Proof of theorem 6.1. (a) The two-sided reflection map can be applied directly, just as in [20, section 5.4]. The limit of the unreflected process is $\mathbf{S}^v - \mathbf{S}^u - \eta\mathbf{e}$, just as in theorem 2.1. For the other cases, we first treat the waiting times. (b) and (c) Let $d_{M_1,t}$ be the M_1 metric on $D([0,t],\mathbb{R})$, as in [20, pp. 82–83]. Let $\|x\|_t \equiv \sup\{|x(s)|\colon 0 \leqslant s \leqslant t\}$, for $t > 0$. Use lemma 6.1 plus the maximum-jump function, as in [20, p. 119], to deduce that

$$d_{M_1,t}\bigl(\mathbf{W}_n^1, \mathbf{W}_n^i\bigr) \leqslant \bigl\|\mathbf{W}_n^1 - \mathbf{W}_n^i\bigr\|_t \Rightarrow 0, \tag{7.34}$$

under the assumption that $P(\mathbf{S}^v \in C) = 1$ for (b) and $P(\mathbf{S}^u \in C) = 1$ for (c). Then use part (a) plus the convergence-together theorem, theorem 11.4.7 of [20]. (d) Use the convergence-together theorem twice more: Use theorem 9.3.2 of [20] to show that the limit $\mathbf{S}_n^v \Rightarrow \mathbf{S}^v$ is unaffected by shifting the index by 1, and note that $V_1/c_n \Rightarrow 0$. Hence

$$d_{M_1,t}\bigl(\mathbf{W}_n^3, \mathbf{W}_n^4\bigr) \Rightarrow 0. \tag{7.35}$$

We now turn to the overflow processes in parts (b)–(d). For parts (b) and (c), we cannot apply lemma 6.1 directly, because we do not have appropriate two-sided bounds on the overflow processes. So, instead, we exploit the one-sided reflection map. First we observe that, since $\mathbf{Q}$ is a random element of D, it almost surely hits the upper barrier at κ after hitting the lower barrier at 0 only finitely often in any finite time interval $[0, t]$. We thus can deduce that the scaled waiting time processes also almost surely hit the upper barrier at K_n/c_n after hitting the lower barrier at 0 only finitely often. We apply the Skorohod representation theorem [20, p. 78], to reduce the argument to a deterministic argument. We consider a single sample path (in a set of probability 1).

We only discuss case (b) because case (c) is essentially the same, and case (d) follows from case (c) by theorem 9.3.2 of [20]. First conside the upper overflow processes in case (b). Note that the difference $L_n^{u,2}(k) - L_n^{u,1}(k)$ can grow only when $W_{n,k}^2 + V_{n,k} > K_n$. We thus consider successive times k at which, first, $W_{n,k}^2 + V_{n,k} > K_n$ and, second, at which $W_{n,k}^2 = 0$. As already observed in our treatment of the waiting times, the difference between these two sequences, $V_{n,k}$, is asymptotically negligible (uniformly for $k \leqslant \lfloor nt \rfloor$). Because of the limits already established for the scaled waiting times, the number of such pairs of hitting times is uniformly bounded in n. We can use the one-sided reflection map with only an upper barrier after any time k for which $W_{n,k}^2 + V_{n,k} > K_n$ until a subsequent time j occurs for which $W_{n,j}^2 = 0$. We can use the one-sided reflection map with only a lower barrier after any time j for which $W_{n,j}^2 = 0$ until a subsequent time k with $W_{n,k}^2 + V_{n,k} > K_n$. The difference $L_n^{u,2}(k) - L_n^{u,1}(k)$ can only change in the subintervals in which the upper barrier is being used. And, with only the upper barrier in effect, the change is bounded by the change in the waiting time difference $|W_{n,k}^2 - W_{n,k}^1|$, which is asymptotically negligible after scaling by virtue of the established limit for the scaled waiting times, under the assumption that $P(\mathbf{S}^v \in C) = 1$. That change can occur during each interval in which the upper barrier is in effect, but since there are only finitely many such intervals, the total difference is asymptotically negliible. Given the result for the upper barrier overflow processes, we obtain the corresponding result for the lower-barrier overflow processes by applying the first line of (6.8). □

8. Conclusion

We have established a heavy-traffic stochastic-process limit for the queue-length and overflow processes in the general $G/G/1/K$ queueing model in section 2. As a consequence (corollary 2.1), we obtain a heavy-traffic limit for the proportion of customers blocked over an initial time interval. If we can justify an interchange of the order of the limits $\lim_{t\to\infty}$ and $\lim_{n\to\infty}$ in (2.18), then that heavy-traffic limit implies a heavy-traffic limit for the associated sequence of steady-state blocking probabilities. Based on that interchange, we obtain the general approximation in (1.6). In the common Brownian case discussed in remark 2.2, we obtain the approximation (2.28). For heavy-tailed service times (with power tails), we obtain the approximations in (3.15) and (3.19). (However, remark 3.1 noted that limits in different orders are inconsistent.)

It is not easy to justify interchanging the order of the two limits in (2.18). We do so for $M/GI/1/K$ and $GI/M/1/K$ special cases in sections 4 and 5 by establishing local heavy-traffic limits, but much more needs to be done.

In the final section, we consider other finite-capacity models. For the waiting times in the $G/G/1/K$ queue with uniformly bounded actual waiting time (model 1), corresponding heavy-traffic stochastic-process limits are easy to establish, because the two-sided reflection map can be applied directly. We show that all the model variants satisfy the same heavy-traffic stochastic-process limits under regularity conditions. It remains to determine what happens when the sample-path-continuity assumptions in theorem 6.1(b), (c) and (d) are not satisfied.

For the $GI/GI/1/K$ versions of these related models, expressions for the steady-state loss proportion are given in (6.21), from which we can develop heuristic heavy-traffic approximations. However, the reliability of these heuristic heavy-traffic approximations is questionable. Paralleling section 4, recent results by Zwart [23] imply corresponding local heavy-traffic limits for the finite dam (model 2) in the $M/GI/1/K$ special case. More generally, it remains to determine when interchanging the order of the two limits $\lim_{t\to\infty}$ and $\lim_{n\to\infty}$ is justified.

References

[1] A.W. Berger and W. Whitt, The Brownian approximation for rate-control throttles and the G/G/1/C queue, J. Discrete Event Dyn. Systems 2 (1992) 7–60.

[2] J. Bertoin, *Lévy Processes* (Cambridge Univ. Press, Cambridge, 1996).

[3] N.H. Bingham, C.M. Goldie and J.L. Teugels, *Regular Variation* (Cambridge Univ. Press, Cambridge, 1987).

[4] A.A. Borovkov, *Stochastic Processes in Queueing Theory* (Springer, New York, 1976).

[5] O.J. Boxma and J.W. Cohen, Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions, Queueing Systems 33 (1999) 177–204.

[6] J.W. Cohen, Some results on regular variation for distributions in queueing and fluctuation theory, J. Appl. Probab. 10 (1973) 343–353.

[7] J.W. Cohen, *The Single-Server Queue* (North-Holland, Amsterdam, 1982).

[8] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1971).

[9] B.V. Gnedenko and A.N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, 2nd ed. (Addison-Wesley, Reading, MA, 1968).

[10] B.V. Gnedenko and I.N. Kovalenko, *Introduction to Queueing Theory* (Israel Program for Scientific Translations, Jerusalem, 1968).

[11] J.M. Harrison, *Brownian Motion and Stochastic Flow Systems* (Wiley, New York, 1985).

[12] P.R. Jelenković, Subexponential loss rates in $GI/GI/1$ queue with applications, Queueing Systems 33 (1999) 91–123.

[13] D. Kennedy, Limit theorems for finite dams, Stochastic Process. Appl. 1 (1973) 269–278.

[14] W. Szczotka, Exponential approximations of waiting time and queue size for queues in heavy traffic, Adv. in Appl. Probab. 22 (1990) 230–240.

[15] W. Szczotka, Tightness of stationary waiting time in heavy traffic, Adv. in Appl. Probab. 31 (1999) 788–794.

[16] H. Takagi, *Queueing Analysis*, Vol. 2: *Finite Systems* (North-Holland, Amsterdam, 1993).

[17] H.C. Tijms, *Stochastic Modelling and Analysis: A Computational Approach* (Wiley, New York, 1986).

[18] W. Whitt, Heavy-traffic approximations for service systems with blocking, AT&T Bell Lab. Tech. J. 63 (1984) 689–708.

[19] W. Whitt, An overview of Brownian and non-Brownian FCLTs for the single-server queue, Queueing Systems 36 (2000) 39–70.

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

[21] W. Whitt, A diffusion approximation for the $G/GI/m/n$ queue, Preprint (2002) (to appear in Oper. Res.).

[22] W. Whitt, Heavy-traffic limits for the $G/H_2^*/n/m$ queue, Preprint (2002).

[23] A.P. Zwart, A fluid queue with a finite buffer and subexponential input, Adv. in Appl. Probab. 32 (2000) 221–243.