Queueing Syst (2009) 63: 13–32
DOI 10.1007/s11134-009-9132-y


INVITED PAPER

# Heavy-traffic extreme value limits for Erlang delay models

**Guodong Pang · Ward Whitt**

Received: 17 March 2009 / Revised: 17 June 2009 / Published online: 5 August 2009


**Abstract** We consider the maximum queue length and the maximum number of idle servers in the classical Erlang delay model and the generalization allowing customer abandonment—the $M/M/n+M$ queue. We use strong approximations to show, under regularity conditions, that properly scaled versions of the maximum queue length and maximum number of idle servers over subintervals $[0,t]$ in the delay models converge jointly to independent random variables with the Gumbel extreme value distribution in the quality-and-efficiency-driven (QED) and ED many-server heavy-traffic limiting regimes as $n$ and $t$ increase to infinity together appropriately; we require that $t_n \to \infty$ and $t_n = o(n^{1/2-\epsilon})$ as $n \to \infty$ for some $\epsilon > 0$.

**Keywords** Erlang models · Many-server queues · Extreme values · Heavy traffic · Diffusion approximations · Strong approximations · Limit theorems

**Mathematics Subject Classification (2000)** 60K25 · 60F05 · 60G70 · 90B22

## 1 Introduction

It is remarkable how persistently the multiserver Erlang $B$ (loss) and $C$ (delay) models have remained the workhorse models for performance analysis of multiserver systems, ever since A.K. Erlang introduced them one hundred years ago. Over the years, the original applications to telecommunication systems have continued, while new applications have emerged, e.g., to new communication systems, call centers, hospitals and other service systems; see [12] for a survey on call centers.

G. Pang · W. Whitt (✉)
IEOR Department, Columbia University, New York, USA
e-mail: ww2040@columbia.edu

G. Pang
e-mail: gp2224@columbia.edu

In this paper, we will once again consider the basic Erlang delay model as well as the Erlang $A$ or Palm $(M/M/n+M)$ model, which includes customer abandonment. The Erlang $B$ and $C$ models appear as the special cases in which the abandonment rate is infinite and zero, respectively. We will be concerned with asymptotic results that facilitate extreme value engineering; see [7]. The idea is to judge whether staffing is appropriate, neither inadequate nor excessive, by looking at the maximum queue length and the maximum number of idle servers over specified time intervals, such as a single hour. Our goal is to apply extreme value theory, as in [11], to develop a systematic way to interpret such extreme value measurements.

However, there are two difficulties, which motivate this research. The first difficulty is discreteness. It is well known that the classical extreme value theory does not apply to integer-valued random variables. For example, with an infinite-server $M/M/\infty$ queue, the steady-state distribution of the number of customers in the system is Poisson, and the maximum of i.i.d. (or weakly dependent) Poisson random variables does not have a nondegenerate extreme value limit; see Example 1.7.14 in [20]. Another example is the $M/M/n/\infty$ queue, which has a steady-state distribution with a geometric upper tail; see Theorem 1.3 of [24].

The second difficulty is the common occurrence of time-varying arrival rates in service systems. The demand typically varies greatly by time of day. In response, the staffing levels typically vary by time of day as well. Since the service times are often short and the arrival rate tends to change relatively slowly, it is often appropriate to use time-varying steady-state performance measures, the pointwise stationary approximation reviewed in [14]. Indeed, we will assume that the subintervals over which we consider extreme values are short enough that the queueing processes can be regarded as approximately stationary; the intervals might be one hour long. However, different hours at different times of the day might have very different arrival rates.

We want a systematic way to relate the extreme values over different hours with very different arrival rates (and staffing). We also want to combine measurements from different hours that may have very different arrival rates. There is a problem, because even with proper staffing set to achieve target service-level constraints throughout the day, the distributions of the maximum queue length and the maximum number of idle servers depend on the arrival rate and the staffing level $n$. We would like performance measures that are easy to interpret directly, without having to relate to the staffing level.

We propose addressing both difficulties for extreme values by applying extreme value approximations (obtained as $t \to \infty$) associated with diffusion approximations (obtained as $n \to \infty$). A key ingredient is appropriate scaling. The diffusion approximations follow from the many-server heavy-traffic limits (as $n \to \infty$) established by Halfin and Whitt [16], Garnett et al. [13] and Whitt [27]. In this limit, properly scaled queueing processes converge to diffusion processes, which have continuous steady-state distributions. In particular, we can then apply extreme value limits for diffusion processes (as $t \to \infty$) established by Davis [10], Borkovec and Klüppelberg [5] and references therein. The scaling in the many-server heavy-traffic limits also allows us to address the difficulty posed by time-varying demand. With the proper scaling, the resulting approximation can be interpreted independent of the staffing level $n$,

provided that the queueing processes can be considered approximately stationary for each $n$.

The procedure we have described involves a two-step limit in which we first let $n \to \infty$, and then afterwards let $t \to \infty$. However, in an application we have fixed $n$ and $t$. We have already observed that we cannot reverse the order of the limits. In order to obtain good approximations for fixed $n$ and $t$, here we consider the *double limit* in which $n \to \infty$ and $t \to \infty$ jointly; i.e., we let $t_n \to \infty$ as $n \to \infty$, imposing a regularity condition that $t_n$ not increase too rapidly, in order to avoid the discreteness problem. After scaling, the resulting approximation will be independent of $n$ and $t$.

Our general approach for obtaining double limits for many-server queues follows the procedure used by Glynn and Whitt [15] to treat single-server queues. As they did, we exploit strong approximations. However, Glynn and Whitt [15] used strong approximations for partial sums by Brownian motion, as in [8]. Since the stochastic process representing the number in system for the $M/M/n+M$ model can be represented in terms of random-time changed Poisson processes, here we will apply strong approximations for Poisson processes by Brownian motion, as in [19]. We also apply the main result from [15] as an intermediate step in our proof; see Lemma 3.2.

Although standard extreme value limits are difficult for the integer-valued queue-length process, there is some relevant literature. Sadowsky and Szpankowski [23] and references therein describe various bounds on the distribution of the maximum queue length for $GI/G/c$ queues. Algorithms have been proposed to compute the distribution of the maximum queue length in a busy period for $M/M/c$ retrial queues for application to call center management by Artalejo et al. [1, 2]. Serfozo [24] and McCormick and Park [21] have obtained extreme value limits for the maximum queue length of $M/M/c$ queues by allowing the birth rates and death rates to vary in a certain manner as the time interval increases. Asmussen [3] has given a good survey on the cycle maxima approach for extreme value limits in queues.

This paper is organized as follows. In Sect. 2, we introduce the scaled processes and state the convergence results. In Sect. 3 we give proofs. We give all details for the proof of Theorem 2.2 and sketch the remaining proofs, which are similar.

## 2 The convergence results

We consider a sequence of $M/M/n+M$ queueing models (with unlimited waiting space) indexed by the number of servers $n$ and let $n \uparrow \infty$. The arrival process is Poisson with rate $\lambda_n$, service times are i.i.d. with an exponential distribution having mean $\mu^{-1}$ and customers abandon independently with an exponential distribution having mean $\theta^{-1}$. For each $n$, we assume that the arrival process, service times and abandonment times are mutually independent. The traffic intensity is $\rho_n \equiv \lambda_n/n\mu$. Assume that $\lambda_n/n \to \lambda \in (0,\infty)$ as $n \to \infty$.

We use the conventional notation: $x \wedge y \equiv \min\{x, y\}$, $x \vee y \equiv \max\{x, y\}$, $x^+ \equiv \max\{x, 0\}$ and $x^- \equiv \max\{-x, 0\}$ for $x, y \in \mathbb{R}$; log is always the natural logarithm (base $e$); $D^k \equiv D([0,\infty), \mathbb{R}^k)$ is the space of right-continuous functions with left limits in $\mathbb{R}^k$, with $D^k \equiv D$ for $k = 1$; $\Rightarrow$ denotes convergence in distribution; see [4] and [26] for background.

### 2.1 The processes of interest

For each $n \geq 1$ and $t \geq 0$, let $X_n(t)$, $Q_n(t) \equiv (X_n(t) - n)^+$ and $I_n(t) \equiv (X_n(t) - n)^-$ represent the number of customers in the system, the queue length, and the number of idle servers, respectively. Let $X_n \equiv \{X_n(t) : t \geq 0\}$, $Q_n \equiv \{Q_n(t) : t \geq 0\}$ and $I_n \equiv \{I_n(t) : t \geq 0\}$ be the associated stochastic processes. Assume that the initial condition $X_n(0)$ is independent of the arrival, service and abandonment processes.

Under those assumptions, the process $X_n$ can be represented as

$$X_n(t) = X_n(0) + A(\lambda_n t) - S\left(\int_0^t \mu\big(X_n(s) \wedge n\big)\, ds\right) - L\left(\int_0^t \theta\big(X_n(s) - n\big)^+ ds\right), \quad t \geq 0, \tag{2.1}$$

where $A \equiv \{A(t) : t \geq 0\}$, $S \equiv \{S(t) : t \geq 0\}$ and $L \equiv \{L(t) : t \geq 0\}$ are mutually independent Poisson processes with unit rate.

Define the running maximum and minimum processes of $X_n$, $M_n \equiv \{M_n(t) : t \geq 0\}$ and $N_n \equiv \{N_n(t) : t \geq 0\}$, respectively, by

$$M_n(t) \equiv \max_{0 \leq s \leq t} X_n(s), \qquad N_n(t) \equiv \min_{0 \leq s \leq t} X_n(s), \quad t \geq 0. \tag{2.2}$$

Define processes $M_n^Q \equiv \{M_n^Q(t) : t \geq 0\}$ and $M_n^I \equiv \{M_n^I(t) : t \geq 0\}$ representing the maximum queue length and the maximum number of idle servers by

$$\begin{aligned} M_n^Q(t) &\equiv \max_{0 \leq s \leq t} Q_n(s) = \big(M_n(t) - n\big)^+, \\ M_n^I(t) &\equiv \max_{0 \leq s \leq t} I_n(s) = \big(n - Nn(t)\big)^+, \quad t \geq 0. \end{aligned} \tag{2.3}$$

We are interested in the asymptotic behavior of $M_n$, $N_n$, $M_n^Q$ and $M_n^I$ as $n \to \infty$ and $t \to \infty$ simultaneously. In the next two subsections, we will state the extreme value limit theorems for these processes in the quality-and-efficiency-driven (QED) and ED regimes. In subsequent subsections we will consider other special cases of the $M/M/n + M$ model: the Erlang $C$ model and the infinite-server queue.

### 2.2 Erlang $A$: QED

With customer abandonment ($0 < \theta < \infty$), in the QED regime the system is *asymptotically critically loaded*; i.e., we assume that

$$\sqrt{n}(1 - \rho_n) \to \beta, \quad \text{as } n \to \infty, \ \beta \in \mathbb{R}. \tag{2.4}$$

The scaling in (2.4) is consistent with the classical square-root staffing principle for large $n$, provided that $\beta > 0$. However, abandonment makes it possible to have $\beta \leq 0$ as well. By assuming (2.4), we are assuming that the system is staffed properly, where the parameter $\beta$ determines the quality of service more precisely.

Define the scaled processes $\bar{X}_n \equiv \{\bar{X}_n(t) : t \geq 0\}$, $\hat{X}_n \equiv \{\hat{X}_n(t) : t \geq 0\}$, $\hat{M}_n \equiv \{\hat{M}_n(t) : t \geq 0\}$, $\hat{M}_n^Q \equiv \{\hat{M}_n^Q(t) : t \geq 0\}$, $\hat{M}_n^I \equiv \{\hat{M}_n^I(t) : t \geq 0\}$, and $\hat{N}_n \equiv \{\hat{N}_n(t) : t \geq 0\}$, where

$$\begin{aligned} \bar{X}_n(t) &\equiv \frac{X_n(t)}{n}, \qquad \hat{X}_n(t) \equiv \frac{X_n(t) - n}{\sqrt{n}}, \\ \hat{M}_n(t) &\equiv \frac{M_n(t) - n}{\sqrt{n}}, \qquad \hat{N}_n(t) \equiv \frac{N_n(t) - n}{\sqrt{n}}, \\ \hat{M}_n^Q(t) &\equiv \frac{M_n^Q(t)}{\sqrt{n}} = \hat{M}_n(t)^+, \qquad \hat{M}_n^I(t) \equiv \frac{M_n^I(t)}{\sqrt{n}} = \big(-\hat{N}_n(t)\big)^+, \quad t \geq 0. \end{aligned} \tag{2.5}$$

It was proved in [13] that, if there exists a random variable $\hat{X}(0)$ such that $\hat{X}_n(0) \Rightarrow \hat{X}(0)$ in $\mathbb{R}$ as $n \to \infty$, then $\hat{X}_n \Rightarrow \hat{X}$ in $D$ as $n \to \infty$, where the limit $\hat{X}$ is the diffusion process with infinitesimal mean $\nu(x) = -\beta\mu - \theta x$ for $x \geq 0$ and $\nu(x) = -\beta\mu - \mu x$ for $x < 0$, and infinitesimal variance $\sigma^2(x) = 2\mu$, i.e.,

$$\begin{aligned} \hat{X}(t) = \hat{X}(0) - \beta\mu t - \int_0^t \mu\big(\hat{X}(s) \wedge 0\big)\, ds \\ - \int_0^t \theta\big(\hat{X}(s) \vee 0\big)\, ds + \sqrt{2\mu} B(t), \quad t \geq 0, \end{aligned} \tag{2.6}$$

where $B \equiv \{B(t) : t \geq 0\}$ is a standard Brownian motion.

Moreover, stationary distributions exist and converge; i.e., $\hat{X}(t) \Rightarrow \hat{X}(\infty)$ and $\hat{X}_n(t) \Rightarrow \hat{X}_n(\infty)$ as $t \to \infty$ for each $n$, and $\hat{X}_n(\infty) \Rightarrow \hat{X}(\infty)$ as $n \to \infty$. Hence, we can initialize with stationary distributions, i.e., we can regard all the processes as stationary processes. That is not required for the extreme value limits, see Theorem 3.1, but it is realistic for applications and clearly should make the approximations perform better for smaller sample size.

Let $\hat{M} \equiv \{\hat{M}(t) : t \geq 0\}$ and $\hat{N} \equiv \{\hat{N}(t) : t \geq 0\}$ be the running maximum and minimum processes of $\hat{X}$, respectively, i.e.,

$$\hat{M}(t) \equiv \max_{0 \leq s \leq t} \hat{X}(s) \quad \text{and} \quad \hat{N}(t) \equiv \min_{0 \leq s \leq t} \hat{X}(s), \quad t \geq 0. \tag{2.7}$$

It follows immediately from applying the continuous mapping theorem [26, Sect. 13.4] that, if there exists a random variable $\hat{X}(0)$ such that $\hat{X}_n(0) \Rightarrow \hat{X}(0)$ in $\mathbb{R}$ as $n \to \infty$, then

$$\big(\hat{M}_n, \hat{N}_n\big) \Rightarrow \big(\hat{M}, \hat{N}\big) \quad \text{in } D^2 \text{ as } n \to \infty. \tag{2.8}$$

We first characterize the extremal behavior of the limit diffusion process $\hat{X}$ in (2.6) in the following proposition. We apply the general extreme value limit theorems established in [5, 10], which are summarized in Sect. 3.1. The proof is given in Sect. 3.2. The extreme value limits for the maximum process $\hat{M}$ and the minimum process $\hat{N}$ are asymptotically independent as $t \to \infty$; see Theorem 3.4 in [10]. In all our extreme value limits, the limiting random variables will have the *standard Gumbel distribution*; let $Z$ denote such a random variable; i.e., $P(Z \leq x) \equiv e^{-e^{-x}}$, $x \in \mathbb{R}$.

The general form of the scaling we obtain in our heavy-traffic extreme value limits combines the heavy-traffic scaling with the extreme value scaling. The extreme value scaling is similar to the scaling for the maximum of i.i.d. random variables with the steady-state distribution of the diffusions process, but there are minor differences. In general, extreme value limits for recurrent diffusion processes are *not* characterized by their steady-state distributions; see Sect. 3.1.

**Proposition 2.1** *The extremal processes $\hat{M}$ and $\hat{N}$ of the limit diffusion process $\hat{X}$ defined in (2.6) and (2.7) have the joint limit*

$$\left(\frac{\hat{M}(t)-b(t)}{a(t)}, \frac{-\hat{N}(t)-d(t)}{c(t)}\right) \quad \Rightarrow \quad (Z_1, Z_2) \quad \text{in } \mathbb{R}^2 \text{ as } t \to \infty, \tag{2.9}$$

*where $Z_1$ and $Z_2$ are independent random variables with the standard Gumbel distribution, and*

$$\begin{aligned}
a(t) &\equiv r/\sqrt{2\log t}, \qquad c(t) \equiv 1/\sqrt{2\log t}, \\
b(t) &\equiv r\sqrt{2\log t} - \beta r^2 + \frac{r}{\sqrt{8\log t}}\left(\log\log t + \log\left(\theta^2\alpha^2\pi^{-1}\left(1-\Phi(r\beta)\right)^{-2}\right)\right), \\
d(t) &\equiv \sqrt{2\log t} - \beta + \left(\frac{\log\log t + \log(\mu^2(1-\alpha)^2\pi^{-1}\Phi(\beta)^{-2})}{\sqrt{8\log t}}\right), \\
\alpha &\equiv \left(1 + \frac{\phi(r\beta)\Phi(\beta)}{r(1-\Phi(r\beta))\phi(\beta)}\right)^{-1}, \quad r \equiv \sqrt{\mu/\theta},
\end{aligned} \tag{2.10}$$

*where $\Phi$ and $\phi$ are the cdf and pdf of the standard normal distribution.*

We remark that the quantity $\alpha$ in Proposition 2.1 plays a key role in the performance measures of Erlang $A$ models; see [13]. Notice that $a(t) \to 0$, as $t \to \infty$, so that we have the limit $\hat{M}(t) - b(t) \Rightarrow 0$ as $t \to \infty$ as a consequence of Proposition 2.1; i.e., there is a concentration about $b(t)$ without additional scaling, and similarly for the other processes. By first letting $n \to \infty$ and then letting $t \to \infty$, we obtain the following extreme value limit theorem for the extremal processes $M_n$, $N_n$, $M_n^Q$ and $M_n^I$.

**Theorem 2.1** *Consider the $M/M/n/\infty + M$ queueing model in the QED regime specified in (2.4). If there exists a random variable $\hat{X}(0)$ such that $\hat{X}_n(0) \Rightarrow \hat{X}(0)$ in $\mathbb{R}$ as $n \to \infty$, then*

$$\begin{aligned}
&\left(\frac{\hat{M}_n(t)-b(t)}{a(t)}, \frac{\hat{M}_n^Q(t)-b(t)}{a(t)}, \frac{-\hat{N}_n(t)-d(t)}{c(t)}, \frac{\hat{M}_n^I(t)-d(t)}{c(t)}\right) \\
&\quad \Rightarrow \quad (Z_1, Z_1, Z_2, Z_2) \quad \text{in } \mathbb{R}^4
\end{aligned} \tag{2.11}$$

*as first $n \to \infty$ and then $t \to \infty$, where $Z_1$ and $Z_2$ are independent with the standard Gumbel distribution and $a(t)$, $b(t)$, $c(t)$ and $d(t)$ are as given in (2.10).*

We next establish an extreme value limit as $n \to \infty$ and $t \to \infty$ simultaneously by imposing a condition that $t_n$ not increase too rapidly.

**Theorem 2.2** *If, in addition to the assumptions of Theorem 2.1, $t_n \to \infty$ and $t_n/n^{1/2-\epsilon} \to 0$ as $n \to \infty$ for some $\epsilon > 0$, then*

$$\left(\frac{\hat{M}_n(t_n) - b_n(t_n)}{a_n(t_n)}, \frac{\hat{M}_n^Q(t_n) - b_n(t_n)}{a_n(t_n)}, \frac{-\hat{N}_n(t_n) - d_n(t_n)}{c_n(t_n)}, \frac{\hat{M}_n^I(t_n) - d_n(t_n)}{c_n(t_n)}\right) \Rightarrow (Z_1, Z_1, Z_2, Z_2), \tag{2.12}$$

*in $\mathbb{R}^4$ as $n \to \infty$, where $Z_1$ and $Z_2$ are independent with the standard Gumbel distribution,*

$$\begin{aligned}
a_n(t_n) &\equiv \frac{r\gamma_n}{\sqrt{2\log t_n}}, \qquad c_n(t_n) = \frac{\gamma_n}{\sqrt{2\log t_n}}, \\
b_n(t_n) &\equiv r\gamma_n\sqrt{2\log t_n} - \beta_n r^2 \\
&\quad + \frac{r\gamma_n}{\sqrt{8\log t_n}}\left(\log\log t_n + \log\left(\theta^2\alpha_n^2\pi^{-1}\left(1 - \Phi(\beta_n r/\gamma_n)\right)^{-2}\right)\right), \\
d_n(t_n) &\equiv \gamma_n\sqrt{2\log t_n} - \beta_n \\
&\quad + \frac{\gamma_n}{\sqrt{8\log t_n}}\left(\log\log t_n + \log\left(\theta^2\left(1-\alpha_n^2\right)\pi^{-1}\left(1 - \Phi(\beta_n/\gamma_n)\right)^{-2}\right)\right), \\
\alpha_n &\equiv \left(1 + \frac{\phi(\beta_n r/\gamma_n)}{r(1-\Phi(\beta_n r/\gamma_n))}\left(\frac{\Phi(\beta_n/\gamma_n)}{\phi(\beta_n/\gamma_n)}\right)\right)^{-1} \to \alpha, \\
\beta_n &\equiv \sqrt{n}(1-\rho_n) \to \beta \quad \text{and} \quad \gamma_n \equiv \sqrt{\left[(\lambda_n/n) + \mu\right]/2\mu} \to 1 \quad \text{as } n \to \infty,
\end{aligned} \tag{2.13}$$

*for $\beta$ in (2.5) and $\alpha$ and $r$ in (2.10). Moreover, the constants $a_n(t_n)$, $b_n(t_n)$, $c_n(t_n)$, and $d_n(t_n)$ can be replaced by $a(t_n)$, $b(t_n)$, $c(t_n)$, and $d(t_n)$, respectively, which are defined in (2.10).*

So far, we have not been able to directly prove the heavy-traffic extreme value limit for the process $\hat{M}_n^I$ in the Erlang $B$ model, but we conjecture that it is given in Theorem 2.2, where we let $\theta \to \infty$ in the normalization constants. Note that $\theta$ does not appear in $c_n$ and only affects $d_n$ through $\alpha_n$, which only appears in the lowest-order term. The $M/M/n + M$ model provides a lower bound.

Based on Theorem 2.2, we can approximate the random vector $(M_n^Q(t), M_n^I(t))$ without scaling in the usual way by

$$\left(M_n^Q(t), M_n^I(t)\right) \approx \left(\sqrt{n}\left[a_n(t)Z_1 + b_n(t)\right], \sqrt{n}\left[c_n(t)Z_2 + d_n(t)\right]\right)$$

for large $t$, where the constants $a_n(t)$, $b_n(t)$, $c_n(t)$ and $d_n(t)$ are given in (2.13), and $(Z_1, Z_2)$ is a pair of independent random variables, each with the Gumbel distribution. Since $E[Z] \approx 0.57721$, the Euler-Mascheroni constant, and $\mathrm{Var}(Z) = \pi^2/6$,

we obtain simple explicit approximations for the mean and variance of the extremal variables $M_n^Q(t)$ and $M_n^I(t)$ for appropriately large $n$ and $t$, e.g.,

$$E\big[M_n^Q(t)\big] \approx \sqrt{n}\big(0.577a_n(t) + b_n(t)\big), \qquad \mathrm{Var}\big(M_n^Q(t)\big) \approx na_n^2(t)\pi^2/6.$$

Moreover, we can also approximate the spread of $X_n$, i.e., $S_n^X(t) \equiv M_n - N_n = M_n^Q(t) + M_n^I(t)$, by

$$S_n^X(t) \approx \sqrt{n}\big(b_n(t) + d_n(t) + a_n(t)Z_1 + c_n(t)Z_2\big)$$

for appropriately large $n$ and $t$.

However, for applications we suggest applying the approximation *with the scaling*. Over different hours, the scaled maximum random variables in (2.12) all approximately have the standard Gumbel distribution independent of $n$, provided that $n$ is not too small. With the scaling, the maximum of $k$ maxima over several separate hours can be approximated as the maximum of $k$ i.i.d. random variables, each with the standard Gumbel distribution, which is again a (nonstandard) Gumbel distribution.

### 2.3 Erlang $A$: ED

In the ED regime, the system is overloaded; i.e., we assume that $\lambda_n = n\lambda$ and $\lambda > \mu$. It is proved in [27] that the fluid scaled process $\bar{X}_n^{ED} \equiv X_n/n$ converges to the constant limit $\bar{X}^{ED}(t) \equiv 1 + (\lambda - \mu)/\theta$ in $D$ as $n \to \infty$ if the scaled initial values converge: $\bar{X}_n^{ED}(0) \Rightarrow \bar{X}^{ED}(0)$ as $n \to \infty$. Define the diffusion scaled processes $\hat{X}_n^{ED} \equiv \{\hat{X}_n^{ED}(t) : t \geq 0\}$, and the scaled extremal processes $\hat{M}_n^{ED} \equiv \{\hat{M}_n^{ED}(t) : t \geq 0\}$ and $\hat{M}_n^{Q,ED} \equiv \{\hat{M}_n^{Q,ED}(t) : t \geq 0\}$ by

$$\begin{aligned}
\hat{X}_n^{ED}(t) &\equiv \frac{X_n^{ED}(t) - n(1 + (\lambda - \mu)/\theta)}{\sqrt{n}} \quad \text{and} \\
\hat{M}_n^{ED}(t) &\equiv \frac{M_n(t) - n(1 + (\lambda - \mu)/\theta)}{\sqrt{n}}, \\
\hat{M}_n^{Q,ED}(t) &\equiv \frac{M_n^Q(t) - n(\lambda - \mu)/\theta}{\sqrt{n}}, \quad t \geq 0.
\end{aligned} \tag{2.14}$$

It is also proved in [27] that if, in addition, $\hat{X}_n^{ED}(0) \Rightarrow \hat{X}^{ED}(0)$ as $n \to \infty$, then $\hat{X}_n^{ED} \Rightarrow \hat{X}^{ED}$ in $D$ as $n \to \infty$, where the limit process $\hat{X}^{ED} \equiv \{\hat{X}^{ED}(t) : t \geq 0\}$ is an Ornstein-Uhlenbeck (OU) process, given by

$$\hat{X}^{ED}(t) = \hat{X}^{ED}(0) + \sqrt{2\lambda}B(t) - \theta \int_0^t \hat{X}^{ED}(s)\,ds, \quad t \geq 0, \tag{2.15}$$

where $B$ is a standard Brownian motion. As in Sect. 2.2, limiting steady-state distributions exist and converge, so that it is natural to assume that we initialize with the steady-state distributions, so that we have stationary processes. The extremal behavior of OU processes has been well studied; see Proposition 3.1. Thus, we have the following extreme value result, paralleling Theorems 2.1 and 2.2.

**Theorem 2.3** *Consider the $M/M/n/\infty + M$ queueing model in the ED regime. If $\hat{X}_n^{ED}(0) \Rightarrow \hat{X}^{ED}(0)$ in $\mathbb{R}$ as $n \to \infty$, then*

$$\left(\frac{\hat{M}_n^{ED}(t) - b(t)}{a(t)}, \frac{\hat{M}_n^{Q,ED}(t) - b(t)}{a(t)}\right) \quad \Rightarrow \quad (Z, Z) \quad \text{in } \mathbb{R}^2, \tag{2.16}$$

*either* (i) *as first $n \to \infty$ and then $t \to \infty$ with*

$$\begin{aligned} a(t) &\equiv \sqrt{\frac{\lambda}{2\theta \log t}}, \\ b(t) &\equiv \sqrt{\frac{2\lambda \log t}{\theta}} + \sqrt{\frac{\lambda}{8\theta \log t}}\big(\log\log t + \log(\theta^2/\pi)\big). \end{aligned} \tag{2.17}$$

*or* (ii) *if in addition $t$ is replaced by $t_n$, where $t_n \to \infty$ and $t_n/n^{1/2-\epsilon} \to 0$ as $n \to \infty$ for some $\epsilon > 0$, where $Z$ is again a random variable with the standard Gumbel distribution.*

## 2.4 Erlang $C$: QED

The story changes if we have no customer abandonment ($\theta = 0$). Without abandonment, the QED regime is again defined by (2.4) but with $\beta > 0$. The representation of the process $X_n$ in (2.1) becomes

$$X_n(t) = X_n(0) + A(\lambda_n t) - S\left(\int_0^t \mu\big(X_n(s) \wedge n\big)\, ds\right), \quad t \geq 0, \tag{2.18}$$

where $A = \{A(t) : t \geq 0\}$ and $S = \{S(t) : t \geq 0\}$ are independent unit-rate Poisson processes. Thus the limit diffusion process $\hat{X}$ in (2.6) becomes

$$\hat{X}(t) = \hat{X}(0) - \beta\mu t - \int_0^t \mu\big(\hat{X}(s) \wedge 0\big)\, ds + \sqrt{2\mu}B(t), \quad t \geq 0, \tag{2.19}$$

where $B \equiv \{B(t) : t \geq 0\}$ is a standard Brownian motion. Halfin and Whitt [16] showed that the steady-state distribution is a combination of a normal pdf below 0 and an exponential pdf above 0; see (3.13) and [6] for an explanation.

Paralleling Proposition 2.1, we have the following characterization of the extremal behavior of the limit process $\hat{X}$ in (2.19).

**Proposition 2.2** *The scaled versions of the extremal processes $\hat{M}$ and $\hat{N}$ of the limit diffusion process $\hat{X}$ defined in* (2.19) *converge jointly*:

$$\left(\frac{\hat{M}(t) - b(t)}{a(t)}, \frac{-\hat{N}(t) - d(t)}{c(t)}\right) \quad \Rightarrow \quad (Z_1, Z_2) \quad \text{in } \mathbb{R}^2 \text{ as } t \to \infty, \tag{2.20}$$

*where $Z_1$ and $Z_2$ are independent with the standard Gumbel distribution, $a(t) \equiv 1/\beta$, $b(t) \equiv (\log t + \log(\beta^2\mu\alpha))/\beta$, $c(t)$ and $d(t)$ are the same as in* (2.10), *where $\alpha$*

*becomes*

$$\alpha \equiv \big(1+\beta\Phi(\beta)/\phi(\beta)\big)^{-1}. \tag{2.21}$$

Thus, paralleling Theorems 2.1 and 2.2, we obtain the following result.

**Theorem 2.4** *Consider the $M/M/n/\infty$ queueing model in the QED regime. If $\hat{X}_n(0) \Rightarrow \hat{X}(0)$ in $\mathbb{R}$ as $n \to \infty$, then (2.12) holds either* (i) *as first $n \to \infty$ and then $t \to \infty$ with the normalization constants $a(t)$, $b(t)$, $c(t)$ and $d(t)$ given in Proposition 2.2, or* (ii) *if in addition $t$ is replaced by $t_n$, where $t_n \to \infty$ and $t_n/n^{1/2-\epsilon} \to 0$ as $n \to \infty$ for some $\epsilon > 0$, with the normalization constants*

$$\begin{aligned} &a_n(t_n) \equiv \gamma_n^2/\beta_n, \qquad b_n(t_n) \equiv \big[\log t_n + \log\big(\mu\beta_n^2\alpha_n/\gamma_n^2\big)\big]\big(\gamma_n^2/\beta_n\big), \\ &\alpha_n \equiv \big(1+(\beta_n/\gamma_n)\Phi(\beta_n/\gamma_n)/\phi(\beta_n/\gamma_n)\big)^{-1}, \end{aligned} \tag{2.22}$$

*$c_n(t_n)$ and $d_n(t_n)$ are given in (2.13) with $\alpha_n$ replaced by $\alpha_n$ in (2.22), and $\beta_n$ and $\gamma_n$ are defined in (2.13). Moreover, the constants $a_n(t_n)$, $b_n(t_n)$, $c_n(t_n)$, and $d_n(t_n)$ can be replaced by $a(t_n)$, $b(t_n)$, $c(t_n)$, and $d(t_n)$, which are defined in Proposition 2.2.*

### 2.5 The infinite-server model

Another important special case of the $M/M/n+M$ model arises with parameter values $\theta = \mu$, which is equivalent to the infinite-server $M/M/\infty$ model. It only requires the limit theorem for $M_n$. The normalization constants in Theorems 2.1 and 2.2 are simplified to

$$\begin{aligned} a(t) &= \frac{1}{\sqrt{2\log t}}, \qquad a_n(t_n) = \frac{\gamma_n}{\sqrt{2\log t_n}}, \\ b(t) &= \sqrt{2\log t} - \beta + \frac{(\log\log t + \log(\mu^2/\pi))}{\sqrt{8\log t}}, \\ b_n(t_n) &= \gamma_n\sqrt{2\log t_n} - \beta_n + \frac{\gamma_n(\log\log t_n + \log(\mu^2/\pi))}{\sqrt{8\log t_n}}, \end{aligned}$$

and $\beta_n$ and $\gamma_n$ are defined in (2.13).

## 3 Proofs

### 3.1 Preliminaries: general results for diffusion processes

The asymptotic behavior of the extremes of general diffusion processes has been established in [5, 10] and references therein. The following is Proposition 3.1 and Corollary 3.2 of [5]. It is significant that, in general, extreme value limits for diffusion processes are *not* determined by the steady-state distribution of the diffusion process, even assuming that it is well defined.

**Theorem 3.1** *Consider the general diffusion process* $\{Y(t) : t \geq 0\}$ *in* $\mathbb{R}$ *defined by*

$$dY(t) = \nu\big(Y(t)\big)\,dt + \sigma\big(Y(t)\big)\,dB(t), \quad t \geq 0,\ Y(0) = y,$$

*and its running maximum process* $M_t^Y \equiv \max_{0\leq s\leq t} Y(s)$. *Suppose that it satisfies the following conditions*: $Y$ *is recurrent, its speed measure* $m$ *has total mass* $|m| < \infty$ *and the scale function* $s$ *satisfies* $s(+\infty) = -s(-\infty) = \infty$. *Then for any* $y \in \mathbb{R}$ *and any* $u_t \uparrow \infty$,

$$\lim_{t\to\infty} \left| P\big(M_t^Y \leq u_t | Y(0) = y\big) - F(u_t)^t \right| = 0,$$

*where* $F$ *is a distribution function, defined by*

$$F(x) = e^{-\frac{1}{|m|s(x)}} \mathbf{1}_{(z,\infty)}(x), \quad x \in \mathbb{R},\ z \in \mathbb{R}, \tag{3.1}$$

*where the values of* $s(x)$ *and* $|m|$ *depend on the choice of* $z$. *Moreover, the tail of* $F$ *satisfies*

$$F^c(x) \equiv 1 - F(x) \sim \left(|m| \int_z^x s'(y)\,dy\right)^{-1} \sim \big(|m|s(x)\big)^{-1} \quad \text{as } x \to \infty.$$

Theorem 3.7 of [5] further characterizes the tail behavior of $F$ in Theorem 3.1 by imposing conditions on the drift coefficient $\nu(x)$ and the volatility coefficient $\sigma(x)$; we only apply part (c) of that theorem. This next result connects the cdf $F$ in (3.1) to the steady-state distribution of the diffusion process under further conditions.

**Theorem 3.2** *Under the conditions of Theorem 3.1, if* $\nu$ *and* $\sigma$ *are differentiable functions on* $(x_0, \infty)$ *for some* $x_0 < \infty$ *such that*

$$\lim_{x\to\infty} \frac{d}{dx}\left(\frac{\sigma^2(x)}{\nu(x)}\right) = 0 \quad \text{and} \quad \lim_{x\to\infty} \frac{\sigma^2(x)}{\nu(x)} \exp\left(-2\int_z^x \frac{\nu(t)}{\sigma^2(t)}\,dt\right) = -\infty,$$

*then* $F^c(x) \sim |\nu(x)|h(x)$ *as* $x \to \infty$, *where* $h$ *is the stationary density of* $Y$.

The maximum domain of attraction of the Gumbel distribution is characterized by a large class of distribution functions; see Theorem 3.3.26 in [11]. An important subclass of such cdf's is the set of von Mises distribution functions $F$, satisfying

$$F^c(x) \equiv 1 - F(x) = c \exp\left\{-\int_z^x \frac{1}{\zeta(t)}\,dt\right\}, \quad z < x < x_F \leq \infty,$$

where $c > 0$ and $\zeta$ is a positive, absolutely continuous function (with respect to Lebesgue measure) with density $\zeta'(x)$ having $\lim_{x\uparrow x_F} \zeta'(s) = 0$. For such a cdf $F$, let $F^{\leftarrow}(p)$ be the inverse, i.e., $x$ is such that $F(x) = p$.

**Theorem 3.3** *If a distribution function* $F$ *is twice differentiable on* $(z, x_F)$ *with positive density* $f$ *and* $F''(x) < 0$ *for* $x \in (z, x_F)$, *then* $F$ *is a von Mises function with*

$\zeta \equiv F^c/f$ *if and only if*

$$\lim_{x\to x_F} \frac{F^c(x)F''(x)}{f^2(x)} = -1.$$

*A von Mises function $F$ belongs to the maximum domain of attraction of the Gumbel distribution and a possible choice of normalization constants is $b_n = F^{\leftarrow}(1-1/n)$ and $a_n = \zeta(b_n)$.*

Extreme value limits for OU processes have been established; see Example 4.1 in [5] for the following result and [9] (also see Theorem 1.9.1 in [8]) for the special case $\alpha = 0$, $\beta = 1$ and $\sigma^2 = 2$ (standard OU process).

**Proposition 3.1** *Let $Y$ be an OU process with drift $v(x) = \alpha - \beta x$ and infinitesimal deviation $\sigma(x) = \sigma > 0$ for $\alpha \in \mathbb{R}$ and $\beta > 0$. Then properly scaled versions of the extremal process $M(t) \equiv \max_{0\le s\le t} Y(s)$, $t \ge 0$, converge*:

$$\frac{M(t) - b(t)}{a(t)} \Rightarrow Z \quad \text{in } \mathbb{R} \text{ as } t \to \infty,$$

*where $Z$ has the standard Gumbel distribution and*

$$a(t) \equiv \frac{\sigma}{2\sqrt{\beta \log t}}, \qquad b(t) \equiv \sqrt{\frac{\sigma^2 \log t}{\beta}} + \frac{\alpha}{\beta} + \frac{\sigma}{4\sqrt{\beta \log t}}\big(\log\log t + \log\big(\beta^2/\pi\big)\big).$$

(We obtain $\log(\beta^2/\pi)$ in the final term of $b(t)$ above instead of $\log(\sigma^2\beta^2/2\pi)$ shown in the bottom line of p. 64 of [5].)

### 3.2 Proofs for $M/M/n/\infty + M$ queues in the QED regime

The limiting diffusion process $\hat{X}$ in (2.6) has the following stationary density; e.g., see [6] or [13]:

$$\begin{aligned} h(x) &= \frac{\alpha\phi(x/r + \beta r)}{r(1 - \Phi(\beta r))}, \quad x \ge 0, \quad \text{and} \\ h(x) &= \frac{(1-\alpha)\phi(x+\beta)}{\Phi(\beta)}, \quad x < 0, \end{aligned} \tag{3.2}$$

for $\alpha$ and $r$ in (2.10).

*Proof of Proposition 2.1* Because of the established asymptotic independence for the extreme value limits of the maximum and the minimum, it suffices to treat them separately. The argument is essentially the same in both cases, so we focus on the maximum. First, it is easy to check that the limit process $\hat{X}$ in the QED regime satisfies the conditions of Theorem 3.2. Hence, we can apply Theorem 3.2 with (3.2) to deduce that

$$F^c(x) \sim |-\mu\beta - \theta x| \frac{\phi(x/r + \beta r)}{r(1 - \Phi(\beta r))}\alpha \quad \text{as } x \to \infty,$$

where $\alpha$ and $r$ are given in (2.10). Then, as $x \to \infty$,

$$F^c(x) \sim \frac{\theta\alpha}{1-\Phi(\beta r)}(x/r+\beta r)^2 \frac{\phi(x/r+\beta r)}{x/r+\beta r} \equiv G^c(x) \sim 1-G(x)$$
$$\sim \frac{\theta\alpha}{1-\Phi(\beta r)}(x/r+\beta r)^2 \Phi^c(x/r+\beta r) \equiv H^c(x).$$

It is well known that $\Phi(\cdot)$ is a von Mises function, from which we deduce that $H$ and $F$ are as well. By Theorem 3.3, we can choose normalization constants

$$b(t) = G^{\leftarrow}(1-1/t) \quad \text{and} \quad a(t) = G^c\big(b(t)\big)/g\big(b(t)\big), \tag{3.3}$$

where

$$g(x) = -\frac{d}{dx}G^c(x) = -\frac{\theta\alpha}{r(1-\Phi(\beta r))}\phi(x/r+\beta r) + (x/r+\beta r)G^c(x).$$

Since $-\log G^c(b(t)) = \log t$, we have

$$-\log\left(\frac{\theta\alpha}{1-\Phi(\beta r)}\right) - \log\big(b(t)/r+\beta r\big) + \frac{1}{2}\log(2\pi) + \frac{1}{2}\big(b(t)/r+\beta r\big)^2 = \log t,$$

and

$$\frac{(b(t)/r+\beta r)^2}{2\log t} \to 1 \quad \text{as } t \to \infty.$$

Hence, we can choose

$$\log\big(b(t)/r+\beta r\big) = \frac{1}{2}(\log 2 + \log\log t) + o(1),$$

so that

$$\begin{aligned}\frac{1}{2}\big(b(t)/r+\beta r\big)^2 &= \log t + \frac{1}{2}(\log 2 + \log\log t) - \frac{1}{2}\log(2\pi) \\ &\quad + \log\left(\frac{\theta\alpha}{1-\Phi(\beta r)}\right) + o(1) \\ &= \log t\left[1 + \frac{1}{2\log t}\Big(\log\log t + \log\big(\theta^2\alpha^2\big(1-\Phi(\beta r)\big)^{-2}\pi^{-1}\big)\Big)\right] \\ &\quad + o(1/\log t).\end{aligned}$$

Thus, we have

$$b(t)/r+\beta r = \sqrt{2\log t}\left[1 + \frac{1}{2\log t}\Big(\log\log t + \log\big(\theta^2\alpha^2\big(1-\Phi(\beta r)\big)^{-2}\pi^{-1}\big)\Big)\right]^{1/2} + o\big(1/\sqrt{\log t}\big)$$

$$= \sqrt{2\log t}\Bigg[1 + \frac{1}{4\log t}\big(\log\log t + \log\big(\theta^2\alpha^2\big(1-\Phi(\beta r)\big)^{-2}\pi^{-1}\big)\big) + o(1/\log t)\Bigg] + o\big(1/\sqrt{\log t}\big),$$

which gives

$$b(t) = r\sqrt{2\log t} - \beta r^2 + \frac{r}{\sqrt{8\log t}}\big(\log\log t + \log\big(\theta^2\alpha^2\pi^{-1}\big(1-\Phi(\beta r)\big)^{-2}\big)\big) + o\big(1/\sqrt{\log t}\big).$$

In addition,

$$a(t) = G^c\big(b(t)\big)/g\big(b(t)\big) \sim \frac{1/t}{(b(t)/r + \beta r)(1/t)} \sim \frac{r}{\sqrt{2\log t}}.$$

For the normalization constants $c(t)$ and $d(t)$, consider the process $Y = -\hat{X}$ with drift $\nu(y) = -\beta\mu - \mu y$ for $y > 0$, and use a similar argument. □

In preparation for our proof of Theorem 2.2, we establish some bounds in the next two lemmas. We will use a strong "sample-path" form of stochastic order for stochastic processes; e.g., see [25]. We write $X_1 \leq_{st} X_2$ for two processes $X_1$ and $X_2$ with sample paths in $D$ if $E[f(X_1)] \leq E[f(X_2)]$ for all nondecreasing measurable real-valued functions $f$ on $D$ for which the expectations are well defined. The following holds by a direct sample-path construction because the birth and death rates are ordered.

**Lemma 3.1** *Under the assumptions of Theorem 2.2, the process $X_n$ in (2.1) can be stochastically bounded above and below*:

$$n - c_2\sqrt{n} - L_n(n\cdot) \leq_{st} X_n(\cdot) \leq_{st} n + c_1\sqrt{n} + U_n(n\cdot), \tag{3.4}$$

*where $U_n(t)$ is the number in system at time $t$ in an $M/M/1$ queue with arrival rate $\lambda_n/n$ and service rate $\theta(1 + c_1/\sqrt{n})$ for $c_1 \in (0,\infty)$ chosen such that $\rho_n^u \equiv (\lambda_n/n)/(\theta(1 + c_1/\sqrt{n})) < 1$ for all $n$ and $\sqrt{n}(1-\rho_n^u) \to \beta_u > 0$ as $n \to \infty$, while $L_n(t)$ is the number in systems in an $M/M/1$ queue with arrival rate $\mu(1 - c_2/\sqrt{n})$ and service rate $\lambda_n/n$ for $c_2 \in (0,\infty)$ chosen such that $\rho_n^l \equiv \mu(1 - c_2/\sqrt{n})/(\lambda_n/n) < 1$ for all $n$ and $\sqrt{n}(1-\rho_n^l) \to \beta_l > 0$ as $n \to \infty$.*

We now apply Lemma 3.1 together with the previous heavy-traffic extreme value limit for single-server queues in [15] to obtain another bound. For any $t > 0$, let $\|x\|_t \equiv \sup_{0\leq s\leq t}\{|x(s)|\}$. Let $\eta$ be the unit constant function, i.e., $\eta(t) \equiv 1$, $t \geq 0$.

**Lemma 3.2** *Under the assumptions of Theorem 2.2, for the given $\epsilon > 0$, $\|\bar{X}_n - \eta\|_{t_n} = o(n^{-(1-\epsilon)/2})$ as $n \to \infty$.*

*Proof* Define $M_n^u(nt_n) \equiv \max_{0\le t\le t_n} U_n(nt)$ and $M_n^l(nt_n) \equiv \max_{0\le t\le t_n} L_n(nt)$. Then by (3.4), we have

$$1 - c_2/\sqrt{n} - M_n^l(nt_n)/n \le_{st} \bar{X}_n(t) \le_{st} 1 + c_1/\sqrt{n} + M_n^u(nt_n)/n, \quad t \in [0, t_n].$$

By the definitions of $U_n$ and $L_n$ in Lemma 3.1, we can apply Theorem 2 and its corollary in [15], and obtain that there exists some large $n_0$ such that for all $n \ge n_0$

$$M_n^l(nt_n) \le O\big(\sqrt{n}\log t_n\big) \quad \text{and} \quad M_n^u(nt_n) \le O\big(\sqrt{n}\log t_n\big).$$

This implies the stated bound. □

*Proof of Theorem 2.2* We start from representation (2.1) and apply the strong approximation for a Poisson process by Brownian motion, see Theorem 2.6.2 in [8] and Lemma 3.1 in [19]. As in Lemma 3.3 of [22], we first apply a crude bound for $X_n(t)$: $X_n(t) \le X_n(0) + A(\lambda_n t)$. By the strong law of large numbers (SLLN) for the Poisson process, we then get $X_n(t) \le c_1 + c_2 nt$ w.p.1 for all $t$, for all $n$ suitably large. Thus, $\int_0^t \bar{X}_n(s)\,ds \le c_1 t_n/n + c_2 t_n^2/2$ w.p.1 for $0 \le t \le t_n$ for all $n$ suitably large. The strong approximation, with that rough bound, allows us to represent the process $X_n$ over the interval $[0, t_n]$ as

$$\begin{aligned} X_n(t) = X_n(0) &+ B_a(\lambda_n t) - B_s\left(n\mu \int_0^t \big(\bar{X}_n(s) \wedge 1\big)\,ds\right) \\ &- B_l\left(n\theta \int_0^t \big(\bar{X}_n(s) - 1\big)^+ ds\right) \\ &+ \lambda_n t - n\mu \int_0^t \big(\bar{X}_n(s) \wedge 1\big)\,ds - n\theta \int_0^t \big(\bar{X}_n(s) - 1\big)^+ ds \\ &+ O\big(\log\big(n\big(t_n^2 \vee 1\big)\big)\big) \end{aligned} \tag{3.5}$$

w.p.1 as $n \to \infty$, where $B_a$, $B_s$ and $B_l$ are mutually independent standard Brownian motions.

By Lemma 3.2, for the given $\epsilon$, there exists some large $n_0$ such that, for all $n \ge n_0$,

$$\left\| \int_0^{\cdot} \big(\bar{X}_n(s) \wedge 1\big)\,ds - \cdot \right\|_{t_n} \le \int_0^{t_n} \big|\bar{X}_n(s) - 1\big|\,ds = o\big(t_n/n^{(1-\epsilon)/2}\big),$$

$$\left| \int_0^{t_n} \big(\bar{X}_n(s) - 1\big)^+ ds \right| = o\big(t_n/n^{(1-\epsilon)/2}\big).$$

(Notice that we are using a "gap" of only $\epsilon/2$ in the exponent of $n$.) We now apply (i) the classical Brownian motion scaling property $\{c^{-1/2}B(ct) : t \ge 0\} \overset{\mathrm{d}}{=} \{B(t) : t \ge 0\}$ for each $c > 0$, where $\overset{\mathrm{d}}{=}$ means equal in distribution on $D$, and (ii) the continuity modulus of standard Brownian motion ($\omega(\delta) = \sqrt{2\delta \log 1/\delta}$ for sufficiently small $\delta$, e.g., see Theorem 2.9.25 in [18]), to obtain a stochastically equivalent alternative representation. In particular, after letting $\hat{Y}_n(0) \equiv \hat{X}_n(0)$ and $p \equiv (1-\epsilon)/2$, we conclude that we can work with a new stochastic process $\hat{Y}_n$ instead of $\hat{X}_n$. In particular,

for each $n \geq n_0$, there are stochastic processes $\tilde{X}_n$ and $\hat{Y}_n$ such that $\tilde{X}_n \stackrel{\mathrm{d}}{=} \hat{X}_n$ for each $n$,

$$\begin{aligned} \left\| \tilde{X}_n - \hat{Y}_n \right\|_{t_n} \equiv \Delta_n = O\left( \left( (t_n/n^p) \log\left(n^p/t_n\right) \right)^{1/2} + \left(\log n t_n^2\right)/\sqrt{n} \right) \quad \text{w.p.1} \quad \text{and} \\ \hat{Y}_n(t) = \hat{Y}_n(0) + B_{a,n}(\lambda_n t/n) - B_{s,n}(\mu t) - \sqrt{n}(1-\rho_n)\mu t \\ - \int_0^t \left[ \mu\left(\hat{Y}_n(s) \wedge 0\right) + \theta \hat{Y}_n(s)^+ \right] ds, \end{aligned} \tag{3.6}$$

where $B_{a,n}$, $B_{s,n}$ and $B_{l,n}$ are mutually independent standard Brownian motions for each $n$ (which in general depend upon $n$ because of our rescaling). Under our assumption that $t_n/n^{(1/2-\epsilon)} \to 0$ as $n \to \infty$, we have $(t_n/n^p)\log(n^p/t_n) \to 0$ and $(\log t_n)/\sqrt{n} \to 0$ as $n \to \infty$, so that $\Delta_n \to 0$ as $n \to \infty$. Thus, we have a stochastically equivalent representation for each $n \geq n_0$ over the interval $[0, t_n]$. Moreover, for showing that the extreme value limits of $\hat{X}_n$ are the same as for $\hat{Y}_n$ (because of the scaling in (2.12)), we exploit the additional relation:

$$(\log n) \left\| \tilde{X}_n - \hat{Y}_n \right\|_{t_n} = (\log n)\Delta_n \to 0 \quad \text{w.p.1} \quad \text{as } n \to \infty, \tag{3.7}$$

which follows easily from the extra $\epsilon/2$ in the exponent of $n$. We use the prefactor $\log n$ to treat the extreme value scaling; note that $a_n(t_n) = O(1/\sqrt{\log t_n}) = o(1)$ as $n \to \infty$; see (3.9)–(3.11).

Now observe that $\hat{Y}_n$ is a diffusion process with infinitesimal drift $\nu_n(x) \equiv -\sqrt{n}(1-\rho_n)\mu - \theta x$ for $x \geq 0$ and $\nu_n(x) = -\sqrt{n}(1-\rho_n)\mu - \mu x$ for $x < 0$ and infinitesimal variance $\sigma_n^2(x) = \lambda_n/n + \mu$. The stationary density of $Y_n$ is given by

$$\begin{aligned} h_{\hat{Y}_n}(x) &= \frac{\phi((x+\beta_n r^2)/(\gamma_n r))}{r\gamma_n(1-\Phi((x+\beta_n r^2)/(\gamma_n r)))}\alpha_n, \quad x \geq 0; \\ h_{\hat{Y}_n}(x) &= \frac{\phi((x+\beta_n)/\gamma_n)}{\gamma_n \Phi(\beta_n/\gamma_n)}(1-\alpha_n), \quad x < 0, \end{aligned}$$

where $\alpha_n$, $\gamma_n$ and $\beta_n$ are given in (2.13), and $r$ is given in (2.10).

Define $\hat{M}_n^Y(t) = \max_{0\leq s\leq t} \hat{Y}_n(s)$ and $\hat{N}_n^Y(t) = \min_{0\leq s\leq t} \hat{Y}_n(s)$. Then by essentially the same argument as in the proof of Proposition 2.1, we obtain

$$\left( \frac{\hat{M}_n^Y(t) - b_n(t)}{a_n(t)}, \frac{-\hat{N}_n^Y(t) - d_n(t)}{c_n(t)} \right) \quad \Rightarrow \quad (Z_1, Z_2) \quad \text{in } \mathbb{R}^2 \text{ as } t \to \infty, \tag{3.8}$$

where $Z_1$ and $Z_2$ are independent with the standard Gumbel distribution, and the normalization constants $a_n(t)$, $b_n(t)$, $c_n(t)$, and $d_n(t)$ are as given in (2.13) with $t_n$ replaced by $t$.

Now we exploit the fact that, for each $n \geq 1$, $\hat{Y}_n$ is a diffusion process just like the limit process $\hat{X}$, with $\nu_n(x) \to \nu(x)$ and $\sigma_n^2(x) \to \sigma^2(x)$ uniformly in $x$. As an immediate consequence, we have $\hat{Y}_n \Rightarrow \hat{X}$ in $D$ as $n \to \infty$. In addition, we can deduce that the extreme value limits for $\hat{Y}_n$ hold as $t \to \infty$ and $n \to \infty$ jointly with $t$

set equal to $t_n$. We establish this step rigorously below, in the final two paragraphs of the proof.

Next, the limit in (3.7), together with the fact that $\tilde{X}_n \stackrel{\mathrm{d}}{=} \hat{X}_n$ for each $n$, implies that the same extreme value limit holds for the scaled versions of $\hat{X}_n$ as $n \to \infty$ with $t_n \to \infty$ at the specified rate. We now give additional details. In particular, we now justify that the scaling functions can be switched in the way claimed. First, it is easy to see from (2.10) and (2.13) that $a_n(t) \to a(t)$, $b_n(t) \to b(t)$, $c_n(t) \to c(t)$, and $d_n(t) \to d(t)$ as $n \to \infty$ for each $t$. For the replacement of $a_n(t_n)$, $b_n(t_n)$, $c_n(t_n)$ and $d_n(t_n)$ by $a(t_n)$, $b(t_n)$, $c(t_n)$ and $d(t_n)$, respectively, in (2.12), some care is needed, because $b(t_n) \to \infty$ and $a(t_n) \to 0$ as $n \to \infty$ (and similarly for $c$ and $d$). We can write

$$\frac{\hat{M}_n(t_n) - b_n(t_n)}{a_n(t_n)} = \frac{\hat{M}_n(t_n) - b(t_n)}{a(t_n)[a_n(t_n)/a(t_n)]} - \frac{b_n(t_n) - b(t_n)}{a(t_n)[a_n(t_n)/a(t_n)]}. \tag{3.9}$$

First, $a_n(t_n)/a(t_n) \to 1$ as $n \to \infty$. Second,

$$\begin{aligned}\frac{b_n(t_n) - b(t_n)}{a(t_n)} &= 2(\gamma_n - 1)\log t_n - r(\beta_n - \beta)\sqrt{2\log t_n} + \frac{1}{2}(\gamma_n - 1)\log\log t_n \\ &\quad + \frac{1}{2}(\gamma_n - 1)\log\left(\theta^2\alpha_n^2\pi^{-1}\left(1 - \Phi(\beta_n r/\gamma_n)\right)^{-2}\right) \\ &\quad - \frac{1}{2}\log\left(\frac{\alpha_n^2(1 - \Phi(\beta_n r/\gamma_n))^{-2}}{\alpha^2(1 - \Phi(\beta r))^{-2}}\right).\end{aligned} \tag{3.10}$$

Note that, by (2.4), we have $\rho_n = 1 - O(1/\sqrt{n})$, and $\gamma_n = \sqrt{(\rho_n + 1)/2} = 1 - O(1/\sqrt{n})$. Hence,

$$|\gamma_n - 1|\log t_n = O\left(\frac{\log t_n}{\sqrt{n}}\right) \to 0 \quad \text{as } n \to \infty. \tag{3.11}$$

Consequently, all the terms in (3.10) are $o(1)$ as $n \to \infty$.

Finally, we justify the joint limit as $n \to \infty$ and $t \to \infty$ in (3.8) where $t = t_n$, satisfying the growth assumption. We do so by bounding the processes $\hat{Y}_n$ above and below by deterministic modifications of the fixed limit process $\hat{X}$. In particular, we establish the strong sample path stochastic ordering

$$(1 + c_n)\hat{X}(t) - d_n^l t \leq_{st} \hat{Y}_n(t) \leq_{st} (1 + c_n)\hat{X}(t) + d_n^u t, \quad t \geq 0, \tag{3.12}$$

where $c_n$, $d_n^l$ and $d_n^u$ are all constants depending on $n$, each being $O(1/\sqrt{n})$. We use the prefactor $(1 + c_n)$ to make the infinitesimal variance match the infinitesimal variance $\sigma_n^2(x) = \sigma_n^2$ of $\hat{Y}_n$. Then we use the stochastic comparison of diffusion processes with common infinitesimal variance but ordered drifts in Theorem 23.5 of [17] to obtain the ordering in (3.12).

The starting point is the elementary observation that, if $Z$ is a diffusion process with infinitesimal mean function $\nu(x)$ and infinitesimal variance $\sigma^2(x)$, and $c$ is a positive constant, then $cZ(t)$ is a diffusion process with infinitesimal mean function $\nu_c(x) = c\nu(x/c)$ and infinitesimal variance function $\sigma_c^2(x) = c^2\sigma^2(x/c)$. That

applies conveniently in our case, because $\sigma^2(x) = \sigma^2$, a constant, while $\nu(x)$ is composed of two linear pieces. Thus, in (3.12) we take $1 + c_n = \sqrt{\sigma_n^2/\sigma^2}$. That yields $(1 + c_n)^2 = \sigma_n^2/\sigma^2 = 1 + O(1/\sqrt{n})$, which implies that $c_n = O(1/\sqrt{n})$. The infinitesimal drift function for $(1 + c_n)\hat{X}$ is $-(1 + c_n)\mu\beta_n - \theta x$ for $x \geq 0$ and $-(1 + c_n)\mu\beta_n - \mu x$ for $x \leq 0$, which differs from the infinitesimal mean function $\nu$ of $\hat{X}$ by a constant function of $x$ depending on $n$. We can now obtain the two bounds in (3.12) by subtracting and adding appropriate functions $d_n t$. These constants $d_n^l$ and $d_n^u$ are both $O(1/\sqrt{n})$ because $c_n = O(1/\sqrt{n})$ and $\beta_n - \beta = O(1/\sqrt{n})$. The proof is completed by observing that the extreme value limits for the bounds, setting $t = t_n$, are the same as for $\hat{X}$ itself, because $t_n/\sqrt{n} \to 0$ as $n \to \infty$ under the assumption on the growth of $t_n$. Hence, the claim (2.12) is proved. □

### 3.3 Sketch of the remaining proofs

*Proof of Theorem 2.3* We can use the same argument as in the proof of Theorem 2.2 except for the following points. First, by the known fluid limit, for any $\epsilon \in (0, (\lambda - \mu)/\theta)$, there exists some $n_1$ such that for all $n \geq n_1$, $\inf_{0 \leq t \leq T} X_n^{ED}(s) \geq n(1 + (\lambda - \mu)/\theta - \epsilon) > n$, for any $T > 0$. So, the strong approximation of $X_n^{ED}$ in (3.5) simplifies:

$$X_n^{ED}(t) = X_n^{ED}(0) + B_a(n\lambda t) - B_s(n\mu t) - B_l\left(n\theta \int_0^t \left(\bar{X}_n^{ED}(s) - 1\right) ds\right)$$
$$+ n\lambda t - n\mu t - n\theta \int_0^t \left(\bar{X}_n^{ED}(s) - 1\right) ds + O\left(\log n\left(t_n^2 \vee 1\right)\right).$$

Second, as in Lemma 3.1, we can stochastically bound $X_n^{ED}$ above and below, but now centering around $n(1 + (\lambda - \mu)/\theta)$ instead of around $n$, by two $M/M/1$ queues to obtain the same bound for $\|\bar{X}_n^{ED} - \eta\|_{t_n}$ as in Lemma 3.2. Third, paralleling (3.6), for each $n \geq n_0$, after letting $\hat{Y}_n(0) \equiv \hat{X}_n^{ED}(0)$ and $p \equiv (1 - \epsilon)/2$, we observe that there are stochastic processes $\tilde{X}_n$ and $\hat{Y}_n$ such that $\tilde{X}_n \stackrel{\mathrm{d}}{=} \hat{X}_n^{ED}$ for each $n$, $\|\tilde{X}_n - \hat{Y}_n\|_{t_n} \equiv \Delta_n$ as in (3.6), and

$$\hat{Y}_n(t) = \hat{Y}_n(0) + B_{a,n}(\lambda t) - B_{s,n}(\mu t) - B_{l,n}\left((\lambda - \mu)t\right) - \theta \int_0^t \hat{Y}_n(s)\, ds.$$

However, now we find that $\hat{Y}_n \stackrel{\mathrm{d}}{=} \hat{X}^{ED}$ for each $n$, where $\hat{X}^{ED}$ is the OU process in (2.15). Hence we can directly apply Proposition 3.1. □

*Proof of Proposition 2.2* We apply the argument used in the proof of Proposition 2.1. First, the stationary density of $\hat{X}$ in (2.19) is given by

$$h(x) = \beta e^{-\beta x}\alpha, \quad x \geq 0 \quad \text{and} \quad h(x) = \frac{\phi(\beta + x)}{\Phi(\beta)}(1 - \alpha), \quad x < 0, \tag{3.13}$$

where $\alpha$ is given in (2.21). Then the tail of the distribution function $F$ in (3.1) becomes

$$F^c(x) \sim \beta^2 \mu\alpha e^{-\beta x} \equiv G^c(x) \sim 1 - G(x) \quad \text{as } x \to \infty.$$

Thus, the constants $a(t)$ and $b(t)$ given in Proposition 2.2 can be obtained by (3.3) where $g(x) = -dG^c(x)/dx = \beta G^c(x)$. Since $-\log G^c(b(t)) = \log t$, we have $-\log(\beta^2\mu\alpha) + \beta b(t) = \log t$. □

*Proof of Theorem 2.4* Again, we can use the same argument as in the proof of Theorem 2.2 with minor modification. First, in Lemma 3.1, we only need to stochastically bound the process $X_n$ from below, which will result in the same bound as given in Lemma 3.2. Second, the stochastically equivalent representation $\hat{Y}_n$ in (3.6) becomes

$$\hat{Y}_n(t) = \hat{Y}_n(0) + B_{a,n}(\lambda_n t/n) - B_{s,n}(\mu t) - \sqrt{n}(1-\rho_n)\mu t - \mu \int_0^t \big(\hat{Y}_n(s) \wedge 0\big)\, ds,$$

with $\hat{Y}_n(0) = \hat{X}_n(0)$. In order to obtain the joint limit as $n \to \infty$ and $t \to \infty$ with $t = t_n$, we can again relate $\hat{Y}_n$ to $\hat{X}$ in the same way. Third, the stationary density of $\hat{Y}_n$ is given by

$$h_{\hat{Y}_n}(x) = \alpha_n \beta_n \gamma_n^{-2} \exp\big(-\beta_n \gamma_n^{-2} x\big), \quad x \geq 0, \quad \text{and}$$

$$h_{\hat{Y}_n}(x) = \frac{\phi((x+\beta_n)/\gamma_n)}{\gamma_n \Phi(\beta_n/\gamma_n)}(1-\alpha_n), \quad x < 0,$$

where $\alpha_n$, $\beta_n$ and $\gamma_n$ are given in (2.22). Then by the argument used to prove Proposition 2.1, we obtain (3.8) where the normalization constants $a_n(t)$, $b_n(t)$, $c_n(t)$ and $d_n(t)$ are given in (2.22) with $t_n$ replaced by $t$. □

**Acknowledgement** This research was supported by NSF grant DMI-0457095.

## References

1. Artalejo, J.R., Economou, A., Gomez-Corral, A.: Applications of maximum queue lengths to call center management. Comput. Oper. Res. **34**, 983–996 (2007)
2. Artalejo, J.R., Economou, A., Lopez-Herrero, M.J.: Algorithmic analysis of the maximum queue length in a busy period for the $M/M/c$ retrial queue. INFORMS J. Comput. **19**, 121–126 (2007)
3. Asmussen, S.: Extreme value theory for queues via cycle maxima. Extremes **1**(2), 137–168 (1998)
4. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley, New York (1999)
5. Borkovec, M., Klüppelberg, C.: Extremal behavior of diffusion models in finance. Extremes **1**, 47–80 (1998)
6. Browne, S., Whitt, W.: Piecewise-linear diffusion processes. In: Dshalalow, J. (ed.) Advances in Queueing, pp. 463–480. CRC Press, Boca Raton (1995)
7. Castillo, E.: Extreme Value Theory in Engineering. Academic Press, New York (1988)
8. Csörgö, M., Révész, P.: Strong Approximations in Probability and Statistics. Akadémiai Kiadó, Budapest (1981)
9. Darling, D.A., Erdös, P.: A limit theorem for the maximum of normalized sums of independent random variables. Duke Math. J. **23**, 143–145 (1956)
10. Davis, R.A.: Maximum and minimum of one-dimensional diffusions. Stoch. Process. Their Appl. **13**, 1–9 (1982)

11. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events for Insurance and Finance. Springer, New York (1997)
12. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review and research prospects. Manuf. Serv. Oper. Manag. **5**, 79–141 (2003)
13. Garnett, O., Mandelbaum, A., Reiman, M.I.: Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. **4**, 208–227 (2002)
14. Green, L.V., Kolesar, P.J., Whitt, W.: Coping with time-varying demand when setting staffing requirements for a service system. Prod. Oper. Manag. **16**, 13–39 (2007)
15. Glynn, P.W., Whitt, W.: Heavy-traffic extreme-value limits for queues. Oper. Res. Lett. **18**, 107–111 (1995)
16. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**, 567–588 (1981)
17. Kallenberg, O.: Foundations of Modern Probability, 2nd edn. Springer, New York (2002)
18. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus, 2nd edn. Springer, Berlin (1991)
19. Kurtz, T.G.: Strong approximation theorems for density dependent Markov chains. Stoch. Process. Their Appl. **6**, 223–240 (1978)
20. Leadbetter, M.R., Lindgren, G., Rootzen, H.: Extremes and Related Properties of Random Sequences and Processes. Springer, New York (1982)
21. McCormick, W.P., Park, Y.S.: Approximating the distribution of the maximum queue length for $M/M/s$ queues. In: Bhat, V.N., Basawa, I.W. (eds.) Queues and Related Models, pp. 240–261. Oxford University Press, London (1992)
22. Pang, G., Talreja, R., Whitt, W.: Martingale proofs of many-server heavy-traffic limits for Markovian queues. Probab. Surv. **4**, 193–267 (2007)
23. Sadowsky, J., Szpankowski, W.: Maximum queue length and waiting time revisited: $GI/G/c$ queue. Probab. Eng. Inf. Sci. **6**, 157–170 (1995)
24. Serfozo, R.F.: Extreme values of birth and death processes and queues. Stoch. Process. Their Appl. **27**, 291–306 (1988)
25. Whitt, W.: Comparing counting processes and queues. Adv. Appl. Probab. **13**, 207–220 (1981)
26. Whitt, W.: Stochastic-Process Limits. Springer, New York (2002)
27. Whitt, W.: Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Manag. Sci. **50**, 1449–1461 (2004)