ELSEVIER

Operations Research Letters 18 (1995) 107–111

operations research letters

# Heavy-traffic extreme-value limits for queues

Peter W. Glynn[a], Ward Whitt[b,*]

[a] Department of Operations Research, Stanford University, Stanford, CA 94305-4022, USA
[b] AT&T Bell Laboratories, Murray Hill, NJ 07974-0636, USA

Received 1 February 1994; revised 1 July 1995

**Abstract**

We consider the maximum waiting time among the first $n$ customers in the GI/G/1 queue. We use strong approximations to prove, under regularity conditions, convergence of the normalized maximum wait to the Gumbel extreme-value distribution when the traffic intensity $\rho$ approaches 1 from below and $n$ approaches infinity at a suitable rate. The normalization depends on the interarrival-time and service-time distributions only through their first two moments, corresponding to the iterated limit in which first $\rho$ approaches 1 and then $n$ approaches infinity. We need $n$ to approach infinity sufficiently fast so that $n(1-\rho)^2 \to \infty$. We also need $n$ to approach infinity sufficiently slowly: If the service time has a $p$th moment for $\rho > 2$, then it suffices for $(1-\rho)n^{1/p}$ to remain bounded; if the service time has a finite moment generating function, then it suffices to have $(1-\rho)\log n \to 0$. This limit can hold even when the normalized maximum waiting time fails to converge to the Gumbel distribution as $n \to \infty$ for each fixed $\rho$. Similar limits hold for the queue-length process.

*Keywords:* Extreme values; Queues; Maximum waiting time; Diffusion approximations; Reflected Brownian motion; Strong approximations; Limit theorems

## 1. Introduction

When doing performance analysis of a service system, we usually try to describe the congestion experienced by a typical arrival, actual or virtual (at an arbitrary time), and thus do the standard steady-state analysis. An alternative approach, known as extreme-value engineering (see [8]), is to describe the maximum congestion experienced over some typical interval. This may still entail steady-state analysis, in that we consider stationary stochastic processes, but now we focus on large values, e.g., the largest customer waiting time over an hour.

In order to be able to effectively use extreme-value engineering in performance analysis, we need to be able to determine the distribution, or at least the mean, of the maximum congestion. This requirement is a major difficulty, because distributions of maximum congestion measures in queueing models are unavailable except in very special cases. However, extreme-value theory comes to our aid. Fundamental limit theorems in extreme-value theory imply that the extreme-value distributions over suitably long intervals can be approximated by a few special distributions [8]. Thus, there is

[*] Corresponding author.

*SSDI* 0167-6377(95)00048-8

statistical regularity associated with looking at extremes, paralleling the more familiar statistical regularity associated with looking at sums and averages, stemming from the central limit theorem.

However, even with the approximating extreme-value distributions, there remains a difficulty, because the parameters of the extreme-value distributions depend upon the queueing processes, and thus the model data, in a complicated way. Berger and Whitt [6] proposed ways to circumvent this difficulty through appropriate approximations for the key parameters. In particular, Berger and Whitt [6] developed and evaluated several heuristic approximations for extreme values of queueing processes over large time intervals. One of these approximations combined the extreme-value limit for reflected Brownian motion (RBM), established in [6], with an approximation of the queueing processes by RBM, which separately can be justified by familiar heavy-traffic limit theorems.

Our purpose here is to establish double limits that determine regions in which the overall approximation in [6] is asymptotically correct. For this purpose, we construct a sequence of queueing system indexed by $n$ in which the traffic intensities $\rho_n$ approach 1 from below as $n \to \infty$. The length of the interval over which the maximum is taken, $t_n$, must also approach infinity, but neither too quickly nor too slowly. We need $(1-\rho_n)^2 t_n \to \infty$ as $n \to \infty$ to have the relevant time in RBM go to infinity, but we also need to impose conditions on how fast $t_n$ grows. These conditions allow the limit to hold even when the normalized maximum wait fails to have the customary extreme-value limit as $t \to \infty$ for fixed $\rho$.

Our principal tools are strong approximation theorems, as in Csörgö and Révész [10]. Strong approximations were used in a similar way to study extreme values associated with sliding window flow control schemes, or scan statistics, in Berger and Whitt [5]. A concrete application of these extreme-value approximations is to compare alternative traffic descriptors in emerging high-speed communication networks; see Berger and Whitt [7].

The specific process we consider is the sequence of waiting times in the GI/G/1 queue (so that $t_n$ above should be an integer), but the argument extends easily to other processes and models, given that corresponding strong approximations hold; e.g., see [11, 17]. The corresponding limit for the discrete queue-length process is interesting because no extreme-value limit holds for each fixed $\rho$. See [12, 14, 16] for background on the basic extreme-value theory. Theorem 2.1 of Horváth [11] provides a strong approximation for the queue length process.

Different heavy-traffic limits for extreme values of queueing processes follow easily from the continuous mapping theorem and related arguments in the cases $(1-\rho_n)\sqrt{t_n} \to c$ or $\rho_n \to \rho > 1$ as $n \to \infty$; see [20, Section 6]. We will review the first case below. The case of $\rho = 1$ is treated in Theorem 9.1 of Iglehart and Whitt [13]. Previous related work on heavy-traffic extreme-value limits for queues has been done by Serfozo [18, 19] and McCormick and Park [15]. See [6] for a numerical evaluation of the approximation through comparisons with simulations.

## 2. Results

For each $n \geqslant 1$, let $W_n \equiv \{W_n(k);\ k \geqslant 0\}$ be a waiting time sequence, defined by $W_n(0) = 0$ and

$$W_n(k+1) = [W_n(k) + \rho_n V_k - U_k]^+, \tag{1}$$

where $[x]^+ = \max\{x, 0\}$, $U \equiv \{U_k\colon\ k \geqslant 1\}$ and $V \equiv \{V_k\colon k \geqslant 0\}$ are independent sequences of i.i.d. nonnegative random variables satisfying

$$EV_k = EU_k = 1, \tag{2}$$

$$\sigma_v^2 \equiv \operatorname{Var} V_k < \infty \quad \text{and} \quad \sigma_u^2 \equiv \operatorname{Var} U_k < \infty, \tag{3}$$

with at least one of $\sigma_v^2 > 0$ and $\sigma_u^2 > 0$. Let

$$A_n = \sum_{k=1}^{n} U_k \quad \text{and} \quad C_n = \sum_{k=0}^{n-1} V_k.$$

Then

$$W_n(k) = S_n(k) - \min_{0 \leqslant j \leqslant k} S_n(j), \tag{5}$$

where

$$S_n(k) = \rho_n C_k - A_k. \tag{6}$$

Let $B \equiv \{B(t): t \geqslant 0\}$ be canonical (drift 0, variance 1) Brownian motion (BM) and let $R \equiv \{R(t): t \geqslant 0\}$ be canonical RBM (with drift $-1$ and variance 1), i.e.,

$$R(t) = B(t) - t - \min_{0 \leqslant s \leqslant t} \{B(s) - s\}, \quad t \geqslant 0. \tag{7}$$

Let

$$M_n(k) = \max_{0 \leqslant j \leqslant k} W_n(j), \quad k \geqslant 0 \tag{8}$$

and

$$M(t) = \max_{0 \leqslant s \leqslant t} R(s), \quad t \geqslant 0. \tag{9}$$

Extreme-value limits for $M_n(k)$ as $k \to \infty$ for any fixed $n$ are given in [12, 16]. These limits require the extra condition

$$E \exp(\varepsilon V_k) < \infty \quad \text{for some } \varepsilon > 0 \tag{10}$$

and more, and involve relatively complicated normalization constants. However, it is natural to expect that the situation should simplify in heavy traffic. To start, we give the standard heavy-traffic result in this setting, paralleling Theorem 9.1 of Iglehart and Whitt [13]. Let $\Rightarrow$ denote convergence in distribution.

**Theorem 1.** *If $\rho_n \uparrow 1$ with $(1 - \rho_n)\sqrt{n} \to c$ as $n \to \infty$, where $0 \leqslant c < \infty$, then*

$$n^{-1/2} M_n(nt) \Rightarrow \left(\frac{\sigma_u^2 + \sigma_v^2}{c}\right) M\left(\frac{c^2 t}{\sigma_u^2 + \sigma_v^2}\right) \quad \text{as } n \to \infty.$$

Theorem 1 is an elementary consequence of the continuous mapping theorem and the basic heavy-traffic limit theorem:

$$n^{-1/2} W_n(\lfloor nt \rfloor) \Rightarrow \left(\frac{\sigma_u^2 + \sigma_v^2}{c}\right) R\left(\frac{c^2 t}{\sigma_u^2 + \sigma_v^2}\right) \quad \text{as } n \to \infty. \tag{11}$$

and Section 6 of Whitt [20]. (See Section 2 of Abate and Whitt [2] for a discussion of scaling time and space.)

Theorem 1 contains the waiting-time part of Theorem 9.1 of [13] as a special case. The distribution of the limit in Theorem 1 can be obtained from the Laplace transform of the associated first-passage-time distribution; see Corollary 3.4.1 of Abate and Whitt [3]. This transform can easily be inverted numerically; see Abate and Whitt [4].

We can combine Theorem 1 above with Theorem 1 of Berger and Whitt [6] to describe the *iterated* limit, as first $\rho \to \infty$ and then $t \to \infty$. For this purpose, let $Z$ be a random variable with the classical Gumbel extreme-value c.d.f., i.e.,

$$P(Z \leqslant x) = \exp(-e^{-z}), \quad -\infty < x < \infty. \tag{12}$$

The following result justifies (4.4) of Berger and Whitt [6] in the case of the waiting times. (Related results hold for queue-length and workload processes.)

**Corollary.** *Under the conditions of Theorem* 1,

$$\left(\frac{2(1 - \rho_n)}{\sigma_u^2 + \sigma_v^2}\right) M_n(nt) - \log\left(\frac{2c^2 t}{\sigma_u^2 + \sigma_v^2}\right) \Rightarrow Z$$

*as first $n \to \infty$ and then $t \to \infty$, where $Z$ is given in* (12).

We now want to generalize the Corollary to Theorem 1 to obtain an appropriate *double limit.* For this, we impose extra moment conditions. There are two cases:

**Theorem 2.** *Suppose that $\rho_n \uparrow 1$ with $(1 - \rho_n)\sqrt{t_n} \to \infty$ as $n \to \infty$. (a) If $EV_k^p < \infty$ for $p > 2$ and $\limsup_{n\to\infty}(1 - \rho_n)t_n^{1/p} < \infty$ as $n \to \infty$, then*

$$\frac{2(1 - \rho_n)M_n(t_n)}{\sigma_u^2 + \sigma_v^2} - \log\left(\frac{2(1 - \rho_n)^2 t_n}{\sigma_u^2 + \sigma_v^2}\right) \Rightarrow Z \quad \text{as } n \to \infty. \tag{13}$$

*(b) If* (10) *holds and $(1 - \rho_n)\log t_n \to 0$ as $n \to \infty$, then* (13) *holds.*

The appeal of Theorem 2 is that is justifies using the Gumbel approximation even when $M_n(k)$ is *not* in its domain of attraction as $k \to \infty$ for each fixed $n$. Dividing by $\log((1 - \rho_n)^2 t_n)$ in Theorem 2, we obtain the following corollary.

**Corollary.** *Under the conditions of Theorem* 2,

$$\frac{2(1 - \rho_n)M_n(t_n)}{\log(t_n)} \Rightarrow \log\left(\frac{2}{\sigma_u^2 + \sigma_v^2}\right) \quad \text{as } n \to \infty.$$

It remains to determine what happens if $t_n \to \infty$ faster than allowed by the conditions of Theorem 2. Iterated limits as first $t \to \infty$ and then $\rho \uparrow 1$ for the decay rate (the Corollary to Theorem 2) are established in Abate et al. [1] and Choudhary and Whitt [9].

## 3. Proof of Theorem 2

We apply the classical KMT (Komlós–Major–Tusnády) strong approximation theorems, as given in Csörgö and Revész [10]. We use Theorem 2.6.3 of [10, p. 107] for part (a) and Theorem 2.6.2 of [10, p. 107] for part (b).

We only display the proof of part (a), because the proof of (b) is almost identical. Under the assumptions of (a), the KMT theorem yields

$$S_n(k) = -(1-\rho_n)k + \sqrt{\rho_n^2\sigma_v^2 + \sigma_u^2}B(k) + \mathrm{o}(k^{1/p}) \quad \text{w.p. } 1 \tag{14}$$

uniformly in $n$. (Under the conditions of (b), the error is $\mathrm{O}(\log k)$.) Given (14),

$$M_n(t_n) = \max_{0 \leqslant k \leqslant t_n} \{-(1-\rho_n)k + \sqrt{\rho_n^2\sigma_v^2 + \sigma_u^2}B(k) - \min_{0 \leqslant j \leqslant k} \{-(1-\rho_n)j + \sqrt{\rho_n^2\sigma_v^2 + \sigma_u^2}B(j)\}\} + \mathrm{o}(t_n^{1/p}) \quad \text{w.p. } 1. \tag{15}$$

Next, by Corollary 1.2.3 of Csörgö and Revész [10, p. 31], we can replace the integer arguments in (15) by continuous ones, i.e.,

$$M_n(t_n) = Y_n(t_n) + \mathrm{o}(t_n^{1/p}) \quad \text{w.p. } 1, \tag{16}$$

where

$$Y_n(t_n) = \max_{0 \leqslant t \leqslant t_n} \{-(1-\rho_n)t + \sqrt{\rho_n^2\sigma_v^2 + \sigma_u^2}B(t) - \min_{0 \leqslant s \leqslant t} \{-(1-\rho_n)s + \sqrt{\rho_n^2\sigma_v^2 + \sigma_u^2}B(s)\}\} + \mathrm{o}(t_n^{1/p}) \quad \text{w.p. } 1. \tag{17}$$

Let $\overset{\mathrm{d}}{=}$ denote equality in distribution. By (16), (17) and the conditions in (a),

$$(1-\rho_n)M_n(t_n) = (1-\rho_n)Y_n(t_n) + \mathrm{o}(1) \quad \text{w.p. } 1, \tag{18}$$

but

$$\begin{aligned} Y_n(t_n) &\overset{\mathrm{d}}{=} \max_{0 \leqslant u \leqslant t_n(1-\rho_n)^2} \{-u + \sqrt{\sigma_u^2 + \rho_n^2\sigma_v^2}B(u) \\ &\quad - \min_{0 \leqslant s \leqslant u} \{-s + \sqrt{\sigma_u^2 + \rho_v^2 + \sigma_u^2}B(s)\}\} \\ &\overset{\mathrm{d}}{=} (\sigma_u^2 + \rho_n^2\sigma_v^2) \max_{0 \leqslant u \leqslant t_n(1-\rho_n)^2} \left\{ -\frac{u}{\sigma_u^2 + \rho_n^2\sigma_v^2} + B\left(\frac{u}{\sigma_u^2 + \rho_n^2\sigma_v^2}\right) \right. \\ &\quad \left. - \min_{0 \leqslant s \leqslant u} \left\{ -\frac{s}{\sigma_u^2 + \rho_n^2\sigma_v^2} + B\left(\frac{s}{\sigma_u^2 + \rho_n^2\sigma_v^2}\right) \right\} \right\} \\ &= (\sigma_u^2 + \rho_n^2\sigma_v^2) M\left(\frac{t_n(1-\rho_n)^2}{\sigma_u^2 + \rho_n^2\sigma_v^2}\right) \end{aligned} \tag{19}$$

for $M$ in (9). By Theorem 1 of Berger and Whitt [6],

$$2M\left(\frac{t_n(1-\rho_n)^2}{\sigma_u^2 + \rho_n^2\sigma_v^2}\right) - \log\left(\frac{2t_n(1-\rho_n)^2}{\sigma_u^2 + \rho_n^2\sigma_v^2}\right) \Rightarrow Z \quad \text{as } n \to \infty \tag{20}$$

for $Z$ in (12). The proof is completed by combining (18)–(20), and noting that $\rho_n \to 1$.

## References

[1] J. Abate, G.L. Choudhury and W. Whitt, "Exponential approximations for tail probabilities in queues, I: waiting times", *Oper. Res.* (1995).

[2] J. Abate and W. Whitt, "Transient behavior of regulated Brownian motion I: starting at the origin", *Adv. in Appl. Probab.* **19**, 560–598 (1987).

[3] J. Abate and W. Whitt, "Transient behavior of the M/M/1 queue via Laplace transforms", *Adv. in Appl. Probab.* **20**, 145–178 (1988).

[4] J. Abate and W. Whitt, "The Fourier-series method for inverting transforms of probability distributions", *Queueing Systems* **10**, 5–88 (1992).

[5] A.W. Berger and W. Whitt, "Asymptotics for open-loop window flow control", *J. Appl. Math. Stochastic Anal.* **7**, 337–356 (1994).

[6] A.W. Berger and W. Whitt, "Maximum values in queueing processes", *Prob. Engr. Inf. Sci.* (1995).

[7] A.W. Berger and W. Whitt, "Comparison of the sliding window and the leaky bucket", *Queueing Systems.* (1995).

[8] E. Castillo, *Extreme Value Theory in Engineering,* Academic Press, New York, 1988.

[9] C.L. Choudhury and W. Whitt, "Heavy-traffic expansions for the asymptotic decay rates in the BMAP/G/1 queue", *Stochastic Models* **10**, 453–498 (1994).

[10] M. Csörgö and P. Révész, *Strong Approximations in Probability and Statistics*, Academic Press, New York, 1981.

[11] L. Horváth, Strong approximations for open queueing networks, *Math. Oper. Res.* **17**, 487–508 (1992).

[12] D.L. Iglehart, "Extreme values in the GI/G/1 queue", *Ann. Math. Statist.* **43**, 627–635 (1972).

[13] D.L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, I", *Adv. in Appl. Probab.* **2**, 150–177 (1970).

[14] M.R. Leadbetter, G. Lindgren and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York, 1983.

[15] W.P. McCormick and Y.S. Park, "Approximating the distribution of the maximum queue length for M/M/s queues", in: U.N. Bhat and I.V. Basawa (eds.), *Queueing and Related Models*, Oxford University Press, 1992, p. 240–261.

[16] A.G. Pakes, "On the tails of waiting-time distributions", *J. Appl. Probab.* **12**, 555–564 (1975).

[17] W. Phillipp and W. Stout, *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*, Mem. Amer. Math. Soc., Vol. 161, Providence, RI, 1975.

[18] R.F. Serfozo, "Extreme values of birth and death processes and queues", *Stochastic Process Appl.* **27**, 291–306 (1988).

[19] R.F. Serfozo, "Extreme values of queue lengths in M/G/1 and GI/M/1 systems", *Math. Oper. Res.* **13**, 349–357 (1988).

[20] W. Whitt, "Some useful functions for functional limit theorems", *Math. Oper. Res.* **5**, 67–85 (1980).