Probability in the Engineering and Informational Sciences, 9, 1995, 375–409. Printed in the U.S.A.

MAXIMUM VALUES IN QUEUEING PROCESSES

ARTHUR W. BERGER
AT&T Bell Laboratories
Holmdel, New Jersey 07733-3030

WARD WHITT
AT&T Bell Laboratories
Murray Hill, New Jersey 07974-0636

Motivated by extreme-value engineering in service systems, we develop and evaluate simple approximations for the distributions of maximum values of queueing processes over large time intervals. We provide approximations for several different processes, such as the waiting times of successive customers, the remaining workload at an arbitrary time, and the queue length at an arbitrary time, in a variety of models. All our approximations are based on extreme-value limit theorems. Our first approach is to approximate the queueing process by one-dimensional reflected Brownian motion (RBM). We then apply the extreme-value limit for RBM, which we derive here. Our second approach starts from exponential asymptotics for the tail of the steady-state distribution. We obtain an approximation by relating the given process to an associated sequence of i.i.d. random variables with the same asymptotic exponential tail. We use estimates of the asymptotic variance of the queueing process to determine an approximate number of variables in this associated i.i.d. sequence. Our third approach is to simplify GI/G/1 extreme-value limiting formulas in Iglehart [25] by approximating the distribution of an idle period by the stationary-excess distribution of an interarrival time. We use simulation to evaluate the quality of these approximations for the maximum workload. From the simulations we obtain a rough estimate of the time when the extreme-value limit theorems begin to yield good approximations.

1. INTRODUCTION AND SUMMARY

This paper is motivated by extreme-value engineering in the performance analysis of service systems. Instead of considering the delay or queue-length distri-

 0269-9648/95 $11.00 + .10

bution in a service system at a particular time, we can consider the distribution of the maximum delay or queue length over a time interval. To use extreme-value engineering effectively in the performance analysis of service systems, we need to be able to describe the distribution of maximum values over relevant time intervals in queueing models of interest. This requirement is a major difficulty, because the exact distributions are unavailable except in very special cases.

We were specifically motivated by a desire to compare two open-loop flow control mechanisms that might be used in emerging high-speed communication networks: the sliding window and the leaky bucket. The leaky bucket can be represented as a G/D/1 queue, hence our interest in the maximum queueing processes. In Berger and Whitt [13] we apply the results here together with results about the sliding window in Berger and Whitt [12] to deduce that the sliding window admits larger bursts than the leaky bucket for given peak rate and given sustainable rate and to quantify the difference. Here we only discuss queues.

To describe extreme-value distributions, it is natural to apply extreme-value limit theorems as in Leadbetter, Lindgren, and Rootzén [29]. Even though the extremes represent unusual behavior for the system, the extreme-value limit theorems show that a certain statistical regularity emerges from considering extremes of stochastic processes (see, e.g., Castillo [16] as well as Leadbetter et al. [29]). However, a queueing model is not an elementary setting for extreme-value limits, because the successive variables in the queueing processes are quite strongly dependent. Nevertheless, extreme-value limit theorems have been proved for queueing processes (see, e.g., Cohen [20], Iglehart [25], Pakes [32], Serfozo [36,37], McCormick and Park [30], Asmussen and Perry [9], Sadowsky and Szpankowski [34,35], Sadowsky [33], and references therein). However, even for relatively simple models such as M/G/1 the exact formulas tend to be somewhat complicated. Moreover, the standard extreme-value limits typically do not even exist for the queue-length processes. (There are bounds and different kinds of limits for the queue-length process, however; see Serfozo [37].) Finally, even when an extreme-value limit theorem applies, it remains to evaluate the quality of the approximation. Most of the previous work on extreme-value limits in queues has not included an examination of the quality of the resulting approximations. (Serfozo [37] is an exception, but he considers a different maximum, in particular, over n busy cycles.)

Our first purpose in this paper is to investigate the quality of the approximations for maximum values in queues provided by the limit theorems when they apply. We find that the extreme-value limits provide excellent approximations for long time intervals (corresponding to thousands or millions of arrivals, which is appropriate for our intended application to communication networks). As part of this investigation, we seek to determine when the time interval is sufficiently long for the limit to become a good approximation. We identify a candidate approximate point where the extreme-value limits begin to kick in, as can be seen from Figures 11–14 (discussed in Section 7). Relative to the remarkably

small number of i.i.d. summands needed to have the normal approximation provided by the central limit theorem perform reasonably well, the length of the interval is quite long however.

Our second purpose is to develop and evaluate relatively simple approximations for the parameters in the limiting extreme-value formulas. Our goal is to obtain simple approximate formulas that are sufficiently accurate for engineering applications. The formulas should capture the essential features of the queueing process and yet not be too complicated. We seek approximations that perform as well for maximum values as previous approximations for steady-state queueing distributions (see, e.g., Whitt [40]). Overall, we regard our quest as a success. We hope that the simple approximations will help facilitate extreme-value engineering.

To be more concrete, let W_n be the waiting time of the nth customer, and let $Q(t)$ be the queue length at time t in a stable queueing model starting with a proper initial distribution, such as empty or in steady state. We are interested in approximate distributions for the associated maximum random variables

$$W_n^* = \max\{W_k : 0 \le k \le n\}, \qquad n \ge 0, \tag{1.1}$$

and

$$Q^*(t) = \max\{Q(s) : 0 \le s \le t\}, \qquad t \ge 0, \tag{1.2}$$

for suitably large values of n and t, respectively.

For the following discussion, let $Q(t)$ be a generic queueing process with associated maximum process $Q^*(t)$. The extreme-value limit theorems suggest that the approximations should be of the form

$$Q^*(t) = \gamma(\log t + \log \beta + Z), \tag{1.3}$$

where t is understood to be relatively large, log is the natural logarithm (base e), Z has the *Gumbel c.d.f.* (classical type-I extreme-value c.d.f.)

$$P(Z \le x) \equiv \Lambda(x) = \exp(-e^{-x}), \qquad -\infty < x < \infty, \tag{1.4}$$

and γ and β are positive constants. For a discrete-time process such as W_n, we would replace t in Eq. (1.3) by n. The specific parameters β and γ in general should depend on the process.

Properties of the Gumbel c.d.f. Λ in Eq. (1.4) are given in Leadbetter et al. [29], Castillo [16], and Chapter 21 of Johnson and Kotz [26]; e.g., $EZ = 0.5772$ (Euler's constant), $\text{Var } Z = \pi^2/6 \approx 1.645$, median $(Z) \approx 0.3667$, and mode$(Z) \approx 0.9624$. As a consequence of Eq. (1.3), we obtain the following approximations for the mean and standard deviation:

$$EQ^*(t) \approx \gamma \log t + \gamma(\log \beta + 0.577) \tag{1.5}$$

and

$$SD(Q^*(t)) \approx 1.28\gamma, \tag{1.6}$$

again for t suitably large. Particularly significant is the form of Eqs. (1.5) and (1.6): *The mean should be linear in* $\log t$, *while the standard deviation should be independent of* t.

Note that $\log \beta + 0.577$ can be negative, so that the approximation for $EQ^*(t)$ in Eq. (1.5) can easily be negative for $t \leq 1$, underscoring the fact that the approximation is only intended for suitably large t. Note that γ is typically the dominant parameter. A candidate approximation for γ is the steady-state mean $EQ(\infty)$. Roughly speaking, Eqs. (1.5) and (1.6) say that $Q^*(t)$ has approximately a mean of $(\log t)EQ(\infty)$ and a standard deviation of $EQ(\infty)$.

We aim to investigate the quality of Eqs. (1.3), (1.5), and (1.6) and develop approximations for the constants γ, β, and $\xi \equiv (\log \beta + 0.577)\gamma$. (We use ξ in addition to β, because Eq. (1.5) can then be rewritten as $\xi + \gamma \log t$; i.e., γ is the slope and ξ is the y intercept for the linear relation in $\log t$.)

We now provide a quick overview of our proposed approximations. (The simulation experiments are described in Section 7.) We present three different approaches. Our *first approach*, yielding the quickest and crudest approximations, follows Whitt [38], where simple heuristic formulas are developed to determine the approximate simulation run lengths required to achieve desired statistical precision in simulations of queueing processes. As in that paper, with our first approach *we specify the class of models and processes we consider by directly assuming that the queueing process can be approximated by one-dimensional reflected (or regulated) Brownian motion* (RBM). (RBM is ordinary Brownian motion with a negative drift and a reflecting barrier at the origin.) See Whitt [38] for additional motivating discussion. For other recent work on Brownian motion approximations for queueing processes, see Asmussen [8], Berger and Whitt [11], and Harrison and Nguyen [24].

To carry out this first approach, we need an extreme-value limit theorem for RBM. Surprisingly, we could not find this result in the literature; hence, we prove it here. Let $\{R(t): t \geq 0\}$ be *canonical* RBM, i.e., RBM with drift coefficient -1 and diffusion coefficient $+1$. Let the associated maximum process be

$$R^*(t) = \sup\{R(s): 0 \leq s \leq t\}, \qquad t \geq 0. \tag{1.7}$$

Let $\Rightarrow$ denote convergence in distribution.

THEOREM 1: *Let* $R(t)$ *be canonical* RBM, *where* $R(0)$ *has a proper initial distribution, and let* Z *have the distribution in Eq.* (1.4). *Then,*

$$2R^*(t) - \log 2t \Rightarrow Z \quad \text{as } t \to \infty.$$

It is easy to see what the statement of Theorem 1 should be by introducing the appropriate scaling of space and time in the known extreme-value limit theorem for the M/M/1 workload process, as we show later in Section 3, but it seems difficult to develop a rigorous proof by this method, because there is an interchange of the limits $t \to \infty$ and $\rho \to 1$, where ρ is the traffic intensity. Hence, we prove Theorem 1 a different way in Section 3.

We mention that the exact distribution for $R^*(t)$ when $R(0) = 0$ is available in the form of a Laplace transform from an early result of Darling and Siegert [21]. A corresponding result for the M/M/1 queue length process is due to Bailey [10]; see Theorem 3.4 and Corollary 3.4.1 of Abate and Whitt [7].

Our *second approach* builds on exact and approximate asymptotic exponential tail behavior for steady-state distributions of queueing processes; see Abate, Choudhury, and Whitt [2,4,5], Asmussen and Perry [9], Choudhury and Lucantoni [17], Choudhury, Lucantoni and Whitt [18], and references therein. For example, suppose that W has the steady-state waiting time distribution. In considerable generality,

$$P(W > x) \sim \alpha e^{-\eta x} \quad \text{as } x \to \infty, \tag{1.8}$$

where η and α are positive constants called the *asymptotic decay rate* and *asymptotic constant*, respectively, and $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$ as $x \to \infty$. The asymptotics in Eq. (1.8) is the starting point for our second approach; i.e., we *assume* that Eq. (1.8) holds. (It is important to note that this need not always be the case, see, e.g., Abate, Choudhury, and Whitt [3] and references therein.)

The key idea in this second approach is that W_n^* should have the same extreme-value limit as

$$X^*_{\lfloor \theta n \rfloor} = \max\{X_k : 0 \le k \le \lfloor \theta n \rfloor\}, \qquad n \ge 0, \tag{1.9}$$

where $\{X_k\}$ is an i.i.d. sequence with X_k distributed the same as W. The parameter θ in Eq. (1.9) is introduced to account for the dependence in the original sequence $\{W_k\}$. The idea is that n dependent random variables should be regarded as approximately equivalent to θn independent random variables. Because the queueing variables tend to be strongly positive correlated except at low traffic intensities, we anticipate that $\theta \ll 1$.

We hasten to point out that the idea of an associated i.i.d. sequence is not new. Indeed, this associated independent sequence is a fundamental notion in extreme-value theory (see Chapter 3 of Leadbetter et al. [29]). In nice situations (without much dependence), the extreme-value limit for a dependent sequence will be identical to the extreme-value limit for the independent sequence (with $\theta = 1$). However, that is not to be expected with queueing processes, because the dependence is quite strong. In particular, condition $D'(u_n)$ on p. 58 of Leadbetter et al. [29] typically does *not* hold.

Under general conditions, which seem hard to verify (see Corollary 3.7.3 of Leadbetter et al. [29]), this second approach is correct for some θ. Indeed, it is consistent with the extreme-value limits for queues (e.g., in Iglehart [25] and Asmussen and Perry [9]). Hence, this approach seems to be a natural heuristic more generally. It leads to tractable formulas, because given Eq. (1.8) the extreme-value limit for the associated independent sequence is easily determined.

It depends on the distribution of W only via the parameters α and η in Eq. (1.8). In particular, the resulting limit is

$$\eta W_n^* - \log(\alpha\theta n) \Rightarrow Z \quad \text{as } n \to \infty, \tag{1.10}$$

where η and α are the asymptotic parameters in Eq. (1.8), θ is the parameter in Eq. (1.9), and Z has the Gumbel c.d.f. in Eq. (1.4). Deducing Eq. (1.10) from Eqs. (1.8) and (1.9) is a standard extreme-value argument. We note that the asymptotic parameters α and η in Eqs. (1.8) and (1.10) are readily computed from transforms by numerical inversion in many cases (see Choudhury and Lucantoni [17]).

Because queue-length processes are integer-valued, we do not quite have Eq. (1.8). Then we have the analog of Eq. (1.8) only as x runs through the integers, which leads to the bounds

$$\alpha \le \underline{\lim}_{x\to\infty} e^{\eta x} P(Q > x) \le \overline{\lim}_{x\to\infty} e^{\eta x} P(Q > x) \le \alpha e^{\eta}. \tag{1.11}$$

The lower bound leads to the analog of Eqs. (1.10) and (1.3) with $\gamma = \eta^{-1}$ and $\beta = \alpha\theta$. For the upper bound we replace β by $\alpha e^{\eta}\theta$, which simply increases $Q^*(t)$ by 1. For integer-valued processes we allow for this error of 1.

To make the second approach work, we need to estimate the parameter θ in Eqs. (1.9) and (1.10). For this purpose we use the *asymptotic variance*

$$\sigma_W^2 \equiv \lim_{n\to\infty} n^{-1} \operatorname{Var}\left(\sum_{k=1}^{n} W_k\right). \tag{1.12}$$

In particular, we estimate θ by

$$\theta \approx \operatorname{Var} W/\sigma_W^2. \tag{1.13}$$

We partly justify Eq. (1.13) by a *cloning heuristic*. We consider the i.i.d. $\{X_k : k \ge 1\}$ and let each variable be repeated (cloned) m times; i.e., we consider the sequence $\{Y_k : k \ge 1\}$ where $Y_{(k-1)m+j} = X_k$ for $j = 1, \ldots, m$. For such sequences obviously $\theta = 1/m$ and $\sigma_Y^2/\operatorname{Var} Y_1 = m$. This is a basis for Eq. (1.13).

Given Eq. (1.13), we need to estimate Var W and σ_W^2. For some models these can be computed (see, e.g., Neuts [31], Whitt [39], and references therein). However, as in Whitt [38], we also suggest using RBM approximations to approximate Var W, σ_W^2, and thus θ. With RBM used to approximate θ, this second approach can be related to the first approach. We find that they support each other, because the resulting formulas are not too different.

Our *third approach* is based on exact extreme value results for the GI/G/1 queue. We simplify exact formulas for the GI/G/1 queue derived by Iglehart [25]. Under conditions equivalent to Eq. (1.8), Iglehart derived the limit of Eq. (1.10) and obtained explicit expressions for the parameters. Here η and α are just as in Eq. (1.8), but θ is the exact value (not based on the heuristic Eq. (1.13)). We develop simple approximations for these parameters. We also consider the GI/G/1 approximation applied to non-GI/G/1 queues.

It turns out that all three approaches lead to approximations of the form of Eq. (1.3). This is to be anticipated, because the steady-state distribution of RBM is exponential, consistent with Eq. (1.8). In the first approach, the key parameter η in Eq. (1.10) is replaced by what turns out to be exactly the first term in the asymptotic expansion for η in powers of $(1-\rho)$ (see Abate et al. [4] and Choudhury and Whitt [19]). In general, we would use the parameter η defined by Eq. (1.8) if it is available and its heavy-traffic approximation if not. The parameter η is often not too difficult to obtain, so that the most difficult part is determining θ in Eq. (1.10). From Eq. (1.10) we see that we actually need to be able to approximate $\log \theta$.

Here is how the rest of this paper is organized. In Section 2 we briefly review the extreme-value limit for the M/M/1 workload process. In Section 3 we use the M/M/1 workload result to develop a heuristic derivation of the RBM extreme-value limit in Theorem 1 and then prove Theorem 1. In Section 4 we describe RBM approximations for generic queueing processes and develop the associated approximations for the queueing maximum processes. In Section 5 we discuss the second "associated i.i.d. sequence" approach further. There we develop the RBM approximation for θ and describe the full approximation. In Section 6 we develop special approximations for the GI/G/1 queue, drawing on the extreme-value limit theorems of Iglehart [25]. A summary of the approximations is given in Table 1. In Section 7 we evaluate the approximations for the special case of the workload by making comparisons with simulations. Finally, in Section 8 we state our conclusions.

2. THE M/M/1 MAXIMUM WORKLOAD PROCESS

Consider an M/M/1 queue with arrival rate ρ and service rate 1, where $0 < \rho < 1$. Let $\{W(t): t \geq 0\}$ be the stationary workload process, i.e., initialized by giving $W(0)$ the steady-state distribution

$$P(W(t) > x) = \rho \exp(-(1-\rho)x), \tag{2.1}$$

for all t. Let the *maximum workload process* be defined by

$$W^*(t) = \sup\{W(s): 0 \leq s \leq t\}, \qquad t > 0. \tag{2.2}$$

Let Z be a random variable with the Gumbel c.d.f. in Eq. (1.4). By Theorem 3 of Iglehart [25] or Cohen [20],

$$(1-\rho)W^*(t) - \log(\rho(1-\rho)^2 t) \Rightarrow Z \quad \text{as } t \to \infty. \tag{2.3}$$

The approximation based on Eq. (2.3) is Eq. (1.3) with $\gamma = (1-\rho)^{-1}$ and $\beta = \rho(1-\rho)^2$.

To obtain Eq. (2.3) from Eq. (9) in Iglehart [25], note that his Eq. (9) is equivalent to

$$\lim_{t \to \infty} P(\gamma W^*(t) - \log(\lambda b^* t/m) \leq x) = \Lambda(x), \tag{2.4}$$

because

$$\Lambda(x + \log c) = \Lambda(x)^{1/c}. \tag{2.5}$$

Then note that $\gamma = 1 - \rho$, $\lambda = \rho$, $m = 1/(1-\rho)$, and $b^* = (1-\rho)$, because $Ee^{\gamma \nu_0} = (1-\gamma)^{-1} = \rho^{-1}$, $Ee^{\gamma S_\alpha} = \rho/(\rho+\gamma) = \rho$, $a(0) = \rho$, and $b(0) = \rho(1-\rho)$.

3. THE MAXIMUM OF CANONICAL RBM

The M/M/1 workload process reveals the basic form of Eq. (1.3) and shows the effect of the traffic intensity ρ. We now consider RBM in order to approximately describe the impact of the variability (possible departure from i.i.d. exponential random variables) in the arrival process and service times.

Hence, let $R(t;\mu,\sigma^2)$ be stationary RBM on the positive real line with drift coefficient μ and diffusion coefficient σ^2, where $\mu < 0$. A stationary version is achieved by letting $R(0;\mu,\sigma^2)$ have the steady-state exponential distribution with mean $\sigma^2/2|\mu|$. Let $R(t) \equiv R(t;-1,1)$ be *stationary canonical* RBM. These processes are related by

$$aR(bt;\mu,\sigma^2) \stackrel{d}{=} R(t;-1,1), \tag{3.1}$$

where $\stackrel{d}{=}$ denotes equality in distribution, $a = |\mu|/\sigma^2$, and $b = \sigma^2/\mu^2$ (see, e.g., Section 2 of Abate and Whitt [6].

Let the *maximum of stationary canonical* RBM be defined as in Eq. (1.7). From Eq. (2.3) and a heavy-traffic limit for $W^*(t)$, we obtain a *heuristic derivation* of Theorem 1. The supporting heavy-traffic limit is

$$\frac{(1-\rho)}{2} W_\rho(2t/(1-\rho)^2) \Rightarrow R(t) \quad \text{as } \rho \to 1, \tag{3.2}$$

where $W_\rho(t)$ indicates the dependence upon ρ (see Section 4 of Whitt [38] for informal discussion and references). By the continuous mapping theorem (in a function space context, as in Billingsley [15]) with the mapping $f(x) = \sup\{x(s) : 0 \le s \le t\}$, we also have

$$\frac{(1-\rho)}{2} W^*_\rho(2t/(1-\rho)^2) \Rightarrow R^*(t) \quad \text{as } \rho \to 1 \tag{3.3}$$

for each t.

Combining Eqs. (2.3) and (3.3), we have

$$R^*(t) \approx \frac{(1-\rho)}{2} W^*_\rho(2t/(1-\rho)^2) \approx \frac{\log(2\rho t) + Z}{2} \approx \frac{\log 2t + Z}{2}, \tag{3.4}$$

which corresponds to Theorem 1. This argument, however, does not yield a proper derivation of Theorem 1 because we have not justified the *interchange of the limits* $\rho \to 1$ and $t \to \infty$. Hence, we give a direct proof for RBM.

PROOF OF THEOREM 1: Just as Iglehart [25] treats the GI/G/1 queue, we break up RBM into contiguous i.i.d. cycles and determine the asymptotic tail behavior within each cycle. We let RBM start at 0 and let the cycles be determined by the first passage from 0 up to 1 and then back down to 0. Let M be the maximum during such a cycle, and let T be the length of a cycle. Clearly, $M > x$ if and only if RBM hits x before it hits 0 starting in 1. By using formula (5) on p. 153 of Kemeny and Snell [27] for simple random walks and a heavy-traffic limit, we see that

$$P(M > x) = \frac{(e^2 - 1)e^{-2x}}{1 - e^{-2x}}, \qquad x \geq 1, \tag{3.5}$$

so that

$$P(M > x) \sim (e^2 - 1)e^{-2x} \quad \text{as } x \to \infty. \tag{3.6}$$

(With Eq. (3.5) there is no interchange of limits.) To do the limiting argument to get Eq. (3.5), let the times between steps be $1/n$, the size of steps be $\pm 1/\sqrt{n}$, and the probability of a step up be $p_n = \left(1 - \left(1/\sqrt{n}\right)\right)/2$ in the nth random walk. This yields canonical RBM in the limit.

By the standard argument (e.g., Lemma 2 of Iglehart [25]) the maximum over n cycles has the limit of Eq. (1.10) with $\eta = 2$ and $\alpha\theta = (e^2 - 1)$. Next we apply Eq. (3.7) of Abate and Whitt [7] to conclude that the expected length of each cycle is

$$ET = (e^2 - 1)/2. \tag{3.7}$$

From renewal theory, if $N(t)$ is the number of cycles in $[0, t]$, then $N(t)/t \to 1/ET$ as $t \to \infty$. Finally, we apply Theorem 3.2 of Berman [14] to treat a random number of cycles, just as in Theorem 2 of Iglehart [25].

So far, we have assumed that RBM starts at 0, but the limit for any other proper initial distribution is the same, because the probability that RBM hits 0 before it hits $x + \log 2t$ approaches 1 as $t \to \infty$. Just condition on whether or not the process hits 0 before $x + \log 2t$. The maximum during the first exceptional cycle is dominated by the maximum over $[0,\infty)$ of ordinary BM (Brownian motion) with drift -1 and this same initial condition. This last maximum is distributed as the initial state plus the independent maximum of BM starting at 0. The maximum of BM starting at 0 is known to have a proper (exponential) distribution.

4. THE RBM APPROXIMATION

As in Whitt [38], we consider a generic stationary queueing process $\{Q_\rho(t) : t \geq 0\}$ indexed by ρ. This might be a queue-length process, a waiting-time process, or something else. We focus on the associated maximum process

$$Q_\rho^*(t) = \sup\{Q_\rho(s) : 0 \leq s \leq t\}, \qquad t \geq 0. \tag{4.1}$$

As our starting point, we *assume* that the RBM approximation

$$(1-\rho)Q_\rho(t/(1-\rho)^2) \approx R(t;a,b) \tag{4.2}$$

is appropriate for some parameters a and b.

The great virtue of Eq. (4.2) is that the complex structure of the queueing process $Q_\rho(t)$ is characterized approximately by the two parameters a and b, together with the traffic intensity ρ. Such approximations tend to perform better when ρ is close to 1; indeed, they often are asymptotically correct as $\rho \to 1$. As discussed in Whitt [38], this RBM approximation is at least roughly appropriate for a large class of single-server queues when ρ is not too small. It is also appropriate for multiserver queues if the number of servers is not too large. For example, for the standard GI/G/m model in which the service time has mean 1, heavy-traffic limit theorems (in which $\rho \to 1$) dictate that for the queue-length process $a = -m$ and $b = m(c_a^2 + c_s^2)$, where c_a^2 and c_s^2 are the squared coefficients of variation (variance divided by the square of the mean) of an interarrival time and service time, respectively (see Section 5.1 of Whitt [38]). If the sequence of interarrival times or service times is not i.i.d., then we would replace c_a^2 and c_s^2 by the corresponding asymptotic variability parameters c_A^2 and c_S^2, i.e., the asymptotic variance (e.g., see (1.12)) divided by the square of the mean.

The same heavy-traffic approximation applies to the continuous-time workload process and associated embedded sequences (the waiting times and queue lengths just before arrivals and just after departures) when the mean service time is 1. The fact that these processes have identical RBM approximations is an indication of the coarseness of the approximation. Of course, further heuristic refinements can be added.

Given Eq. (4.2), Eq. (3.1), and Theorem 1, we obtain the associated approximations

$$Q_\rho(t) \approx \frac{b}{|a|(1-\rho)} R(a^2(1-\rho)^2 t/b) \tag{4.3}$$

and

$$\begin{aligned} Q_\rho^*(t) &\approx \left[\frac{b}{|a|(1-\rho)}\right] R^*(a^2(1-\rho)^2 t/b) \\ &\approx \left[\frac{b}{2|a|(1-\rho)}\right] (\log(2a^2(1-\rho)^2 t/b) + Z), \end{aligned} \tag{4.4}$$

where a and b are the drift and diffusion parameters in initial RBM approximation (4.2) and Z has the Gumbel c.d.f. in Eq. (1.4). Note that Eq. (4.4) can be expressed in the same form as Eq. (1.3) by writing the log term as a sum of two log terms. In particular, the parameters for form (1.3) are

$$\gamma = \frac{b}{2|a|(1-\rho)} \quad \text{and} \quad \beta = \frac{2a^2(1-\rho)^2}{b}. \tag{4.5}$$

We remark that γ in Eq. (4.5) coincides with the RBM approximation for $EQ_\rho(t)$ in Eq. (4.3) if we use $ER(a^2(1-\rho)^2 t/b) \approx ER(\infty) = \frac{1}{2}$. This explains the simple approximation $\gamma \approx EQ_\rho(t)$ in Section 1.

From Choudhury and Whitt [19], we know that, in considerable generality, $2|a|(1-\rho)/b$ is the heavy-traffic approximation, i.e., the first term in an asymptotic expansion in powers of $(1-\rho)$, for the asymptotic decay rate η in the analog of Eq. (1.8). We would always use η in Eq. (1.8) if it is available. We regard $2|a|(1-\rho)/b$ as a convenient approximation.

It is also important to note that, mathematically, for Eq. (4.2) we rely only on the limit $\rho \to 1$, whereas for Eq. (4.4) we rely on the two limits $\rho \to 1$ and $t \to \infty$. The asymptotic correctness of formula (4.4) requires not only that ρ be suitably close to 1 but also that ρ approach 1 in proper relation to t as $t \to \infty$. As in Whitt [38] and Asmussen [8], we argue that the RBM time scaling indicates that we should relate time t to $(1-\rho)^{-2}$.

Formally, we can do this by defining a family of models indexed by ρ. For the RBM approximation to be meaningful, t should be of order $(1-\rho)^{-2}$. Hence, we can proceed as follows. First, we choose t^* suitably large for the RBM extreme-value limit provided by Theorem 1 to yield a good approximation to $R^*(t)$ for $t \geq t^*$. Then, we let t in model ρ be

$$\hat{t}(\rho) = \frac{b\hat{t}}{a^2(1-\rho)^2}, \tag{4.6}$$

where $\hat{t} \geq t^*$. Then, under regularity conditions, we will have

$$(1-\rho)Q_\rho^*(\hat{t}(\rho)) \Rightarrow \frac{b}{|a|} R^*(\hat{t}) \quad \text{as } \rho \to 1, \tag{4.7}$$

so that, for suitably high ρ and this $\hat{t}(\rho)$.

$$Q_\rho^*(\hat{t}(\rho)) \approx \frac{b}{|a|(1-\rho)} R^*(\hat{t}) \tag{4.8}$$

and Eq. (4.4) should be good approximations.

If $t(\rho)$ is much smaller than Eq. (4.6), then the RBM extreme-value limit may not yield a good approximation. On the other hand, if $t(\rho)$ is much larger than Eq. (4.6), then different extreme-value behavior for the queueing process with fixed ρ may dominate. Whether or not $t(\rho)$ growing faster than Eq. (4.6) will cause a problem no doubt depends on the model. Choudhury and Whitt [19] show that for the asymptotic decay rate η the limits $t \to \infty$ and $\rho \to 1$ can be interchanged when the steady-state distribution has an exponential tail. Having $t(\rho)$ as in Eq. (4.6) will produce the exponential tail in the double limit even when it is not present as $t \to \infty$ for fixed ρ. See Glynn and Whitt [22] for additional discussion.

We now see how RBM approximation (4.4) applies to the M/M/1 workload process discussed in Section 2. By Eq. (3.1), Eq. (3.2) is equivalent to

$$(1-\rho)W_\rho(t/(1-\rho)^2) \approx R(t;-1,2). \tag{4.9}$$

Hence, Eq. (4.9) satisfies Eq. (4.2) with $a=-1$ and $b=2$. We thus can apply Eq. (4.4) to get

$$W_\rho^*(t) \approx \left(\frac{1}{1-\rho}\right)(\log((1-\rho)^2 t) + Z), \tag{4.10}$$

which agrees with Eq. (2.3) asymptotically as $\rho \to 1$. As in Section 4.4 of Whitt [38], we can use Eqs. (2.3) and (4.10) to develop an M/M/1 *refinement* to Eq. (4.4); i.e., we insert a ρ inside the logarithm in Eq. (4.4) to obtain

$$Q_\rho^*(t) \approx \frac{b}{2|a|(1-\rho)}(\log(2a^2\rho(1-\rho)^2 t/b) + Z). \tag{4.11}$$

We regard Eq. (4.11) as our *refined* RBM *approximation* for the maximum of the queueing process $Q_\rho(t)$. For example, for the queue length, workload, and waiting time processes in the GI/G/1 queue (with mean service time 1), we would use Eq. (4.11) with $a=-1$ and $b=c_a^2+c_s^2$. By the preceding discussion, our M/M/1 refinement makes this formula exact for the M/M/1 workload process for all ρ, $0<\rho<1$. Even though the M/M/1 queue-length process does not have an extreme-value limit of form (2.3), we obtain approximations for it from Eq. (4.11) as well. As in Eqs. (1.5) and (1.6), Eq. (4.11) immediately yields associated approximations for the mean and standard deviation of $Q_\rho^*(t)$, based on properties of Z stated in Section 2, namely,

$$EQ_\rho^*(t) \approx \frac{b}{2|a|(1-\rho)}(\log t + \log(2a^2\rho(1-\rho)^2/b) + 0.577) \tag{4.12}$$

and

$$SD\, Q_\rho^*(t) \approx \frac{0.64b}{|a|(1-\rho)}. \tag{4.13}$$

5. THE ASSOCIATED i.i.d. SEQUENCE APPROXIMATION

In this section we complete the development of the approximation based on an associated i.i.d. sequence begun in Section 1. As in Section 4, we start with a generic stationary queueing process $Q_\rho(t)$ and consider the associated maximum process $Q_\rho^*(t)$ in Eq. (4.1). Our key assumption, as in Eq. (1.8), is that $Q_\rho(0)$ has an exponential tail; i.e.,

$$P(Q_\rho(0) > x) \sim \alpha e^{-\eta x} \quad \text{as } x \to \infty, \tag{5.1}$$

where x runs through the integers if $Q_\rho(0)$ is integer-valued. Now the different processes in the same model need not have the same parameters. (However,

quite a bit is known about the relations among the asymptotic parameters of the standard queueing processes; see, e.g., Abate et al. [2,5].)

Given Eq. (5.1), we reason as in Section 1 (thinking of continuous-valued processes) and let our approximation be

$$Q_\rho^*(t) \approx \frac{\log(\alpha\theta t) + Z}{\eta}, \tag{5.2}$$

where α and η come from Eq. (5.1), Z has the Gumbel c.d.f. in Eq. (1.4), and the parameter θ is approximated by

$$\theta \approx \text{Var } Q/\sigma_Q^2, \tag{5.3}$$

where Var Q is the variance of the steady-state variable $Q_\rho(0)$ and σ_Q^2 is the asymptotic variance, i.e., Eq. (1.12) for discrete-time processes and

$$\sigma_Q^2 = \lim_{t\to\infty} t^{-1} \text{Var}\left(\int_0^t Q_\rho(s)\, ds\right) \tag{5.4}$$

for continuous-time processes.

If we can calculate η, α, σ_Q^2, and Var Q, then we are done. If not, then we resort to further approximation, depending on what is still needed. If we do not know any of these four parameters, then we would rely on the RBM approximation in Eq. (4.11). The RBM approximation also yields approximations for the individual parameters. In particular, given RBM approximation (4.2), we let η coincide with γ in Eq. (4.5) and $\alpha = 1$. Given Eq. (4.3) and the fact that the steady-state distribution of RBM is exponential, we would approximate Var $Q_\rho(0)$ by

$$\text{Var } Q_\rho(0) \approx (EQ_\rho(0))^2 \approx \frac{b^2}{4a^2(1-\rho)^2}. \tag{5.5}$$

Finally, given Eq. (4.2), we would approximate the asymptotic variance by using the known asymptotic variance of RBM, i.e., by

$$\sigma_Q^2 \approx b^3/2a^4(1-\rho)^4, \tag{5.6}$$

just as in Eq. (36) of Whitt [38]. From Eqs. (5.3), (5.5), and (5.6), we obtain the approximation

$$\theta \approx a^2(1-\rho)^2/2b. \tag{5.7}$$

Combining Eq. (5.2) with all these individual RBM approximations, we obtain the RBM approximation in Eq. (4.11) except that the argument in the logarithm is $\alpha\theta t \approx a^2(1-\rho)^2 t/2b$ instead of $2a^2\rho(1-\rho)^2 t/b$, i.e., the argument there is $4\rho/\alpha$ times the argument here. Because $\rho = \alpha$ for M/M/1, these approximations can be made consistent by modifying our approximation for θ in Eq. (5.3), i.e., by replacing Eq. (5.3) with

$$\theta \approx 4 \text{ Var } Q/\sigma_Q^2, \tag{5.8}$$

and we make this heuristic refinement. Note that this changes the numerator of Eq. (5.2) by $\log 4 \approx 1.38$. If t is suitably large, then this change will not be too great relatively. The fact that the two approximation methods yield similar results lends support to both of them.

In summary, the associated i.i.d. sequence approximation plus additional RBM approximations (5.5) and (5.6) and heuristic approximation (5.8) yield the simple approximation

$$\theta = 2a^2(1-\rho)^2/b. \tag{5.9}$$

In many cases, it will be appropriate to approximate α in Eqs. (5.1) and (5.2) by 1, but we note that is not always so. In a queue with an arrival process that is the superposition of many independent non-Poisson processes, the asymptotic constant α can be far from 1 (see Choudhury et al. [18]). Abate et al. [4] propose the approximation

$$\alpha = \eta E Q_\rho(0), \tag{5.10}$$

which is useful if approximations are already available for η and the mean.

6. APPROXIMATIONS FOR THE GI/G/1 QUEUE

Consider a GI/G/1 model with i.i.d. service times independent of i.i.d. interarrival times. Let U be a generic interarrival time having Laplace-Stieltjes transform (LST) $\hat{f}_U(s) \equiv Ee^{-sU}$ and mean ρ^{-1}, and let V be a generic service time having LST $\hat{f}_V(s) \equiv Ee^{-sV}$ and mean 1. Let W and L be the steady-state waiting time and workload, respectively. For a large class of GI/G/1 queues, Iglehart [25] proved that limit (1.10) for the waiting times and the workload is correct, where the parameters η and α are the asymptotic decay rate and asymptotic constant in the tail asymptotics of Eq. (1.8). The asymptotic decay rate η is the same for the waiting time and workload. The asymptotic constants are related by

$$\alpha_L = \frac{\alpha_W \rho(Ee^{\eta V} - 1)}{\eta}, \tag{6.1}$$

where the subscript W (L) indicates waiting time (workload) (see Theorem 2 of Abate et al. [5]).

For the GI/G/1 queue, the asymptotic decay rate η is the root of the equation

$$Ee^{s(V-U)} \equiv \hat{f}_V(-s)\hat{f}_U(s) = 1. \tag{6.2}$$

The key condition is that such a root exists. We also require that the distribution of $V-U$ be nonlattice and that $0 < E[(V-U)e^{\eta(V-U)}] < \infty$. Algorithms for computing η, α_W, and α_L in GI/G/1 and BMAP/G/1 queues are described in Abate et al. [1,2,4] and Choudhury and Lucantoni [17]; we use them.

From Theorems 2 and 3 of Iglehart [25], we see that the remaining parameter θ in Eq. (1.10) is

$$\theta_W = P(W = 0)(1 - Ee^{-\eta I}) \tag{6.3}$$

for the waiting times and

$$\theta_L = \rho P(W = 0) Ee^{\eta V}(1 - Ee^{-\eta I}) \tag{6.4}$$

for the workload, where in both cases $P(W = 0)$ is the steady-state probability that an arrival finds an idle server, which is the reciprocal of the mean busy cycle. In both cases I is an idle period.

For M/G/1 the idle periods have the same distribution as the exponential interarrival times. Hence, for M/G/1

$$1 - Ee^{-\eta U} = \frac{\eta}{\rho + \eta}. \tag{6.5}$$

For the M/G/1 queue we also have $P(W = 0) = 1 - \rho$. Hence, for M/G/1,

$$\theta_W = (1 - \rho)\frac{\eta}{\rho + \eta} \tag{6.6}$$

for the waiting times and

$$\theta_L = \rho(1 - \rho) Ee^{\eta V}\frac{\eta}{\rho + \eta} \tag{6.7}$$

for the workload. As shown in Section 2, $\alpha_L \theta_L = \rho(1 - \rho)^2$ for M/M/1.

Following Halfin [23], we suggest approximating the distribution of I more generally (within GI/G/1) by the stationary-excess (or equilibrium residual life) distribution associated with U, i.e., with LST $\rho(1 - \hat{f}_U(s))/s$. Thus, for GI/G/1 we obtain the approximation

$$(1 - Ee^{-\eta I}) \approx 1 - \frac{\rho(1 - \hat{f}_U(\eta))}{\eta}. \tag{6.8}$$

For the M/G/1 queue $\hat{f}_U(s) = \rho/(\rho + s)$, so that $\rho(1 - \hat{f}_U(s))/s = \hat{f}_U(s)$, as it should.

Following Kraemer and Langenbach-Belz [28], we approximate the steady-state delay probability $P(W > 0)$ by

$$P(W > 0) = \rho + (c_a^2 - 1)\rho(1 - \rho)h(\rho, c_a^2, c_s^2), \tag{6.9}$$

where

$$h(\rho, c_a^2, c_s^2) = \begin{cases} \dfrac{1 + c_a^2 + \rho c_s^2}{1 + \rho(c_s^2 - 1) + \rho^2(4c_a^2 + c_s^2)}, & c_a^2 \le 1 \\ \dfrac{4\rho}{c_a^2 + \rho^2(4c_a^2 + c_s^2)}, & c_a^2 > 1, \end{cases} \tag{6.10}$$

and as before c_a^2 and c_s^2 are the squared coefficients of variation of an interarrival time and a service time. (An algorithm for the exact value of $P(W > 0)$ is also given in Abate et al. [1].) Using Eqs. (6.9) and (6.10), we obtain concrete approximation formulas for the GI/G/1 waiting times and workload in terms of the transform values $\hat{f}_U(\eta), \hat{f}_V(-\eta)$, the asymptotic parameters η and α_N and α_L, and the basic parameters ρ, c_a^2, and c_s^2.

Following Abate et al. [4], we can develop further approximations for the transform values $\hat{f}_V(\eta)$ and $\hat{f}_V(-\eta)$ by expanding in Taylor series; i.e.,

$$\hat{f}_V(-\eta) = 1 + \eta + \eta^2 \frac{(c_s^2 + 1)}{2} + o(\eta^2) \quad \text{as } \eta \to 0 \tag{6.11}$$

and

$$\hat{f}_U(\eta) = 1 - \frac{\eta}{\rho} + \eta^2 \frac{(c_a^2 + 1)}{2\rho^2} + o(\eta^2) \quad \text{as } \eta \to 0. \tag{6.12}$$

Combining Eqs. (6.9) and (6.12), we obtain

$$(1 - Ee^{-\eta I}) \approx \frac{\eta(c_a^2 + 1)}{2\rho}. \tag{6.13}$$

Abate et al. [4] also develop approximations for the waiting-time parameters η and α_W, e.g., $\alpha_W \approx \eta EW$. The first two terms of an asymptotic expansion for η for waiting times in GI/G/1 are

$$\eta = \frac{2(1-\rho)}{c_a^2 + c_s^2}\,(1 - (1-\rho)\eta^* + O((1-\rho)^2) \quad \text{as } \rho \to 1, \tag{6.14}$$

where

$$\eta^* = \frac{(2\nu_3 - 3c_s^2(c_s^2 + 2)) - (2u_3 - 3c_a^2(c_a^2 + 2))}{3(c_a^2 + c_s^2)^2} \tag{6.15}$$

with $\nu_3 = E[V^3]$ and $u_3 = E[(\rho U)^3]$ (see Theorem 3 of Abate et al. [4] and Choudhury and Whitt [19].

In summary, the exact GI/G/1 formulas for the waiting time and workload are Eq. (1.10) with Eqs. (6.3) and (6.4), where the two asymptotic constants are related by Eq. (6.1). We then approximate $P(W = 0)$ by Eq. (6.9) and $(1 - Ee^{-\eta I})$ by Eqs. (6.8) and (6.13). This produces an approximation for θ that depends only on the parameters η, ρ, c_a^2, and c_s^2. We give an approximation for η in Eq. (6.14) as well that depends on ρ, c_a^2, and c_s^2 and the third-moment parameters u_3 and ν_3.

Finally, we indicate how we can apply the GI/G/1 approximation in this section to non-GI/G/1 queues. We assume that Eq. (1.10) still applies. We start with the asymptotics in Eq. (1.8), assuming that the asymptotic decay rate η and the asymptotic constant α are available. For example, these asymptotic parameters are available for BMAP/G/1 queues [2] and other models whenever the

transform of the steady-state distribution is available [17]. Then our first approximation is to act as if Eqs. (6.3) and (6.4) are valid. A simple approach is to apply Eqs. (6.9) and (6.13), either directly or with c_a^2 and c_s^2 replaced by the asymptotic variability parameters c_A^2 and c_S^2 (e.g., c_A^2 is the asymptotic variance of the interarrival times divided by the square of the mean). As an alternative to Eq. (6.9), we often can calculate $P(W = 0)$. (This is so for BMAP/G/1 models. For multiserver GI/G/m queues, an approximation is given in Whitt [40].) For the idle time we can use Eq. (6.8) or (6.13).

7. SIMULATION EXPERIMENTS

In this section we describe simulation experiments conducted, first, to determine whether or not linear relations (1.5) and (1.6) tend to be valid for some parameters γ and β and, second, to evaluate the quality of the proposed approximations. For this purpose we used a simulation program written in C and run on a SUN SPARC-2 workstation.

In the basic experiment we consider the workload process in seven different single-server queueing models (M/M/1, M/D/1, M/H_2/1, H_2/D/1, H_2/H_2/1, MMPP/D/1, and MMPP/H_2/1) each for two traffic intensities ($\rho = 0.7$ and $\rho = 0.9$). In all cases the mean service time is $EV = 1$.

The H_2 (hyperexponential) distribution is the mixture of two exponential distributions, i.e., with density function

$$f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \quad x \geq 0, \tag{7.1}$$

with balanced means ($p/\lambda_1 = (1-p)/\lambda_2$) and squared coefficient of variation $c^2 = 4$. The third parameter is determined by the mean (see, e.g., Abate and Whitt [6, p. 592]).

The Markov-modulated Poisson process (MMPP) is an example of a nonrenewal arrival process. The MMPP is a Poisson process in which the rate itself evolves as a continuous-time Markov chain. The state space of this underlying Markov chain is called the environment. Extreme-value limit theorems for M, H_2 and MMPP arrival processes with phase-type service were established by Asmussen and Perry [9]. Our MMPP has two environment states with the mean holding time in each being 10. The arrival rates in the two states are 1.6ρ and 0.4ρ, respectively. Hence, for the traffic intensities we consider, the instantaneous traffic intensity (the arrival rate in that environment state) in one state exceeds 1. For traffic intensity $\rho = 0.7$, the squared coefficient of variation of a single stationary interarrival time is $c_a^2 = 1.44$, whereas the asymptotic variability parameter (the asymptotic variance of the interarrival times, as in Eq. (1.12), divided by the square of the mean) is $c_A^2 = 3.52$. The fact that $c_A^2 > c_a^2$ reflects the nonrenewal property, i.e., c_A^2 includes all the autocovariance terms. For $\rho = 0.9$ these parameters are $c_a^2 = 1.48$ and $c_A^2 = 4.24$.

For each case we performed 20 independent replications of a simulation of duration $2^{14} \times 10^3 = 16{,}384{,}000$ time units and recorded the maximum work-

load in the queue at the 15 epochs $2^k \times 10^3$ for $k = 0,1,\ldots,14$. (The expected number of arrivals in each run is thus about 16.4ρ million.) We then calculated the sample means and sample standard deviations of the 20 data points for each of the 15 time points. We fit regression lines (i.e., by least squares) to these sample means and sample standard deviations. We crudely estimate the statistical precision of slope and intercept estimates by using standard regression formulas, which assume that the errors are i.i.d. Clearly, the errors at successive times are not independent here, so that our estimates of 95% confidence intervals for the slopes and intercepts are only rough approximations, which may significantly underestimate the true variability.

Figures 1 and 2 display simulation results for the sample means and the sample standard deviations in the M/M/1 queue. The 15 time values are displayed along with 95% confidence intervals and a regression line in each case. (For the individual time points, the confidence intervals are approximately valid, because the random variables at single time points are independent. However, they are likely to have approximately the Gumbel distribution in Eq. (1.4) rather than the assumed t-distribution.) In Figure 1 we also display one of the 20 individual sample paths. The linear relation in Eq. (1.5) is not so evident from a single run, because of the random fluctuations on individual sample paths. However, the linear relations in Eqs. (1.5) and (1.6) are clear when we look at the data from 20 independent runs.

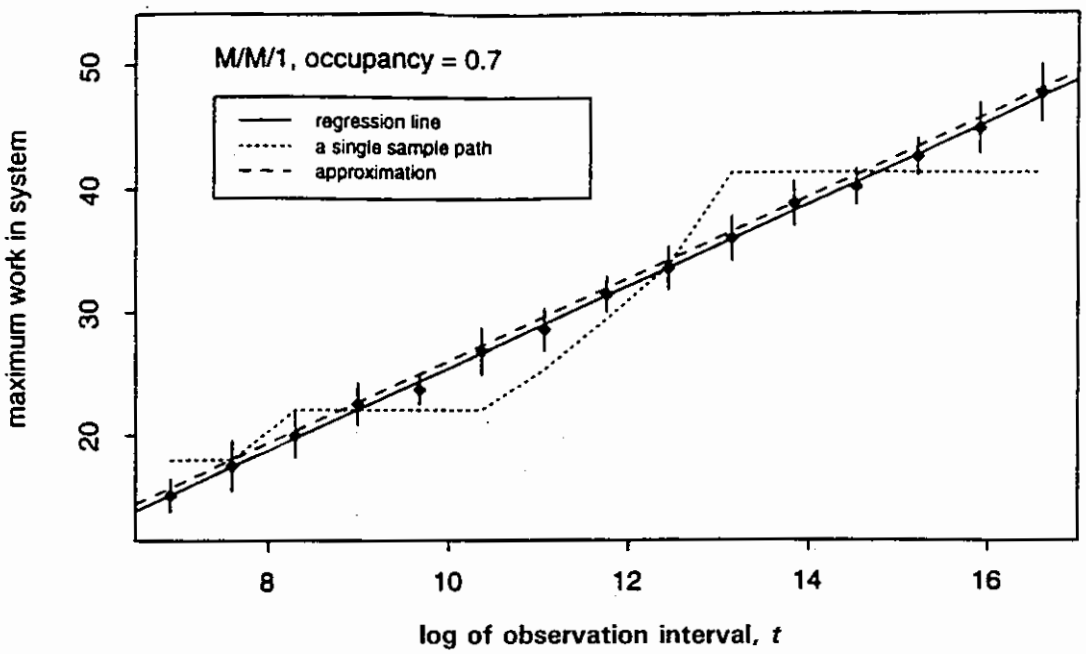

FIGURE 1. Mean of maximum work in system realized over 20 sample paths.

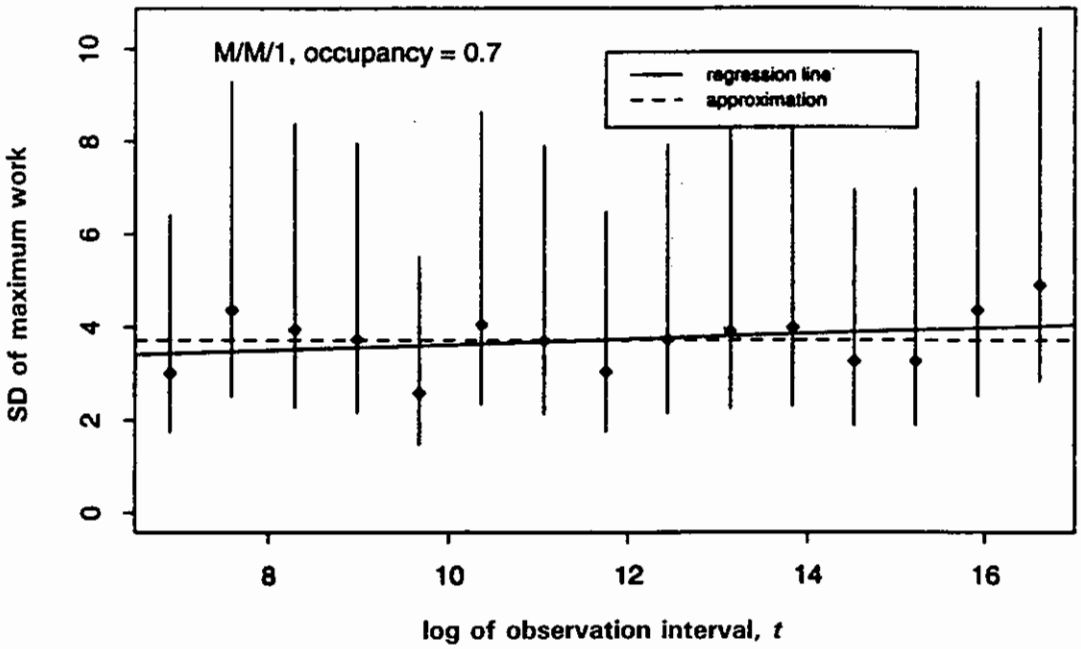

FIGURE 2. Standard deviation of maximum work in system realized over 20 sample paths.

Figures 1 and 2 also show approximations. Only one approximation is shown in each case because all three approximations coincide with the approximation provided by the extreme-value limit theorem in the M/M/1 special case.

We now consider more general models. The different approximations are summarized in Table 1. Figures 3–10 show both the simulations and the approximations for the mean of the maximum workload as a function of time in 8 of the 14 cases. The six cases omitted in Figures 3–10 are the ones for which the approximation performs best. The figures in these six cases show strong agreement, essentially the same as for the $H_2/H_2/1$ queue with $\rho = 0.7$ shown in Figure 6. In all 14 cases the linear relation in Eq. (1.5) is clearly present. For all 14 cases, the simulation estimates and approximations for the slope and intercept are given in Tables 2 and 3, respectively. As background, we also give the values of the asymptotic constant α (for the workload analog of Eq. (1.8)) in Table 2.

In this experiment the associated i.i.d. sequence approximation for the mean performs well in all cases. The GI/G/1 approximation is also excellent for all GI/G/1 queues, but its extension to non-GI/G/1 queues does not perform so well for the MMPP arrival process. The RBM approximation is excellent for some models, but not all. In particular, the RBM approximation for the slope γ degrades dramatically when the service-time distribution is deterministic (M/D/1, H_2/D/1, and MMPP/D/1). (The deterministic service-time distribution is known to be a difficult case for heavy-traffic approximations for GI/G/m queues; see Berger and Whitt [11] and Whitt [40].) Consistent with

Table 1. Summary of the Parameter Approximations for the Expected Maximum Workload in Eq. (1.5)

Method	Slope γ	β
Crude mean	EL	-0.577
Refined RBM (4.11)	$\dfrac{b}{2\lvert a\rvert(1-\rho)}$	$\dfrac{2a^2\rho(1-\rho)^2}{b}$
Associated i.i.d. sequence with Eq. (5.9)	$1/\eta$ for decay rate η in Eq. (1.8)	$\alpha\theta$ for α in Eq. (1.8) $\theta \approx 4$ Var L/σ_L^2 $\approx 2a^2(1-\rho)^2/b$
GI/G/1	$1/\eta$ for decay rate η in Eq. (1.8)	$\alpha\theta_L$ for α in Eq. (1.8) $\theta_L = \rho P(W=0)Ee^{\eta V}(1-Ee^{-\eta I})$
GI/G/1 approximation	$1/\eta$ for decay rate η in Eq. (1.8)	$\alpha\theta_L$ for α in Eq. (1.8) $\theta_L = \rho P(W=0)Ee^{\eta V}(1-Ee^{-\eta I})$ $P(W=0)$ from Eq. (6.9) $(1-Ee^{-\eta I})$ from Eq. (6.8) or (6.13)
GI/G/1 approximation for non-GI/G/1	$1/\eta$ for decay rate η in Eq. (1.8)	GI/G/1 approximation plus Eq. (6.13) with c_A^2 instead of c_a^2

intuition, the RBM approximation for the slope improves as the traffic intensity increases. Hence, with D service, the RBM approximation for the slope γ is not bad at $\rho = 0.9$ but rather poor at $\rho = 0.7$. Nevertheless, the simple RBM approximation for γ may serve as a useful rough approximation, because it is much easier to calculate than the exact asymptotic decay rate η in Eq. (1.8). The experiments here give a good idea about the accuracy to expect.

Even in cases where RBM's rough approximation for the slope is not sufficient, the RBM approximation for $\log \beta$ may still be useful. In particular, the approximation obtained by combining the reciprocal of the asymptotic decay rate with the RBM approximation for $\log \beta$ is very close to the associated i.i.d. sequence approximation with Eq. (5.9), as could be anticipated by comparing Eqs. (4.11) and (5.9). (Both are based at least in part on RBM.)

Because the GI/G/1 extreme-value formulas do not apply directly to non-GI/G/1 queues, we investigated the GI/G/1 approximation to the MMPP/G/1 queues more carefully. The specific GI/G/1 approximation shown in Table 3 and Figures 8–10 is Eq. (1.10) with the exact asymptotic parameters α and η computed using the program described in Abate et al. [2] plus Eqs. (6.4), (6.9), and (6.13) with c_a^2 replaced by the asymptotic variability parameter c_A^2 in both

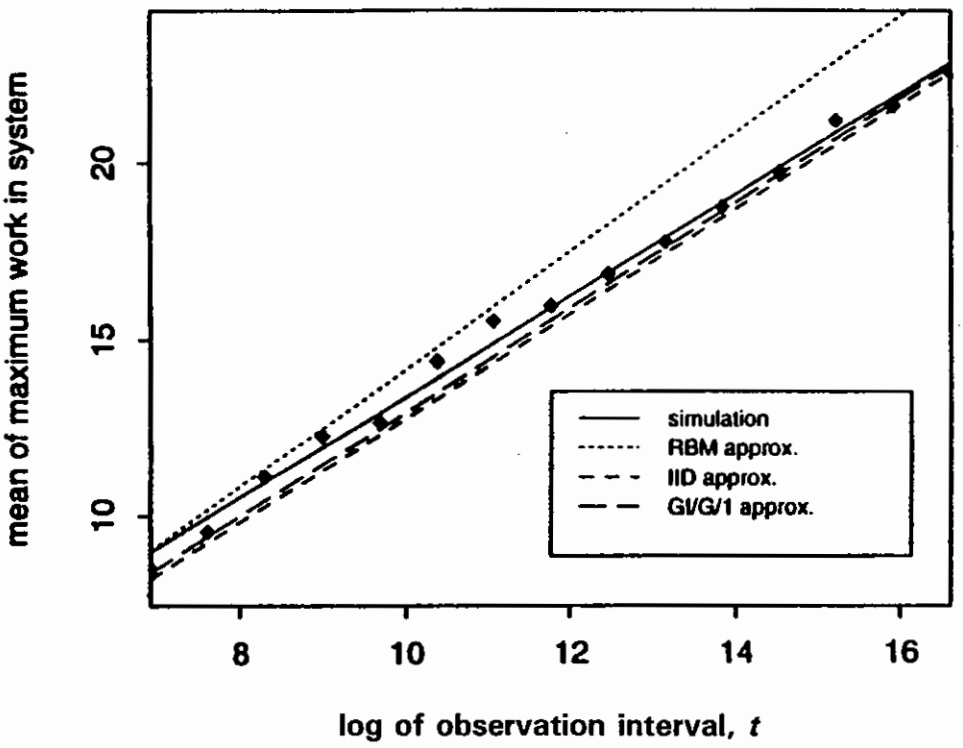

FIGURE 3. M/D/1, $\rho = 0.7$.

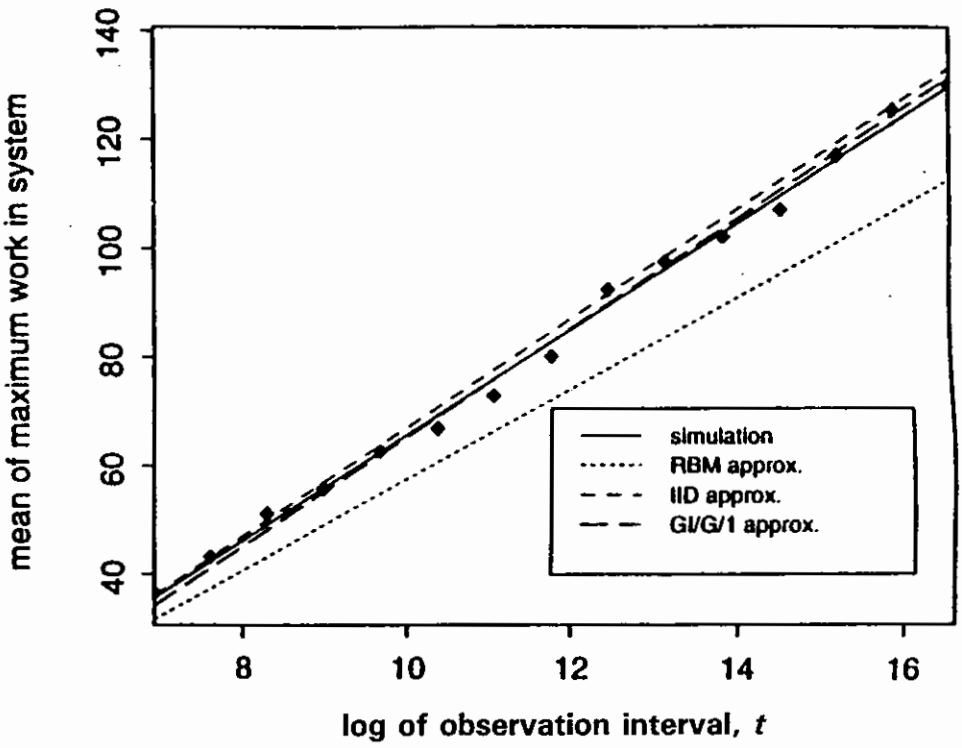

FIGURE 4. M/H_2/1, $\rho = 0.7$.

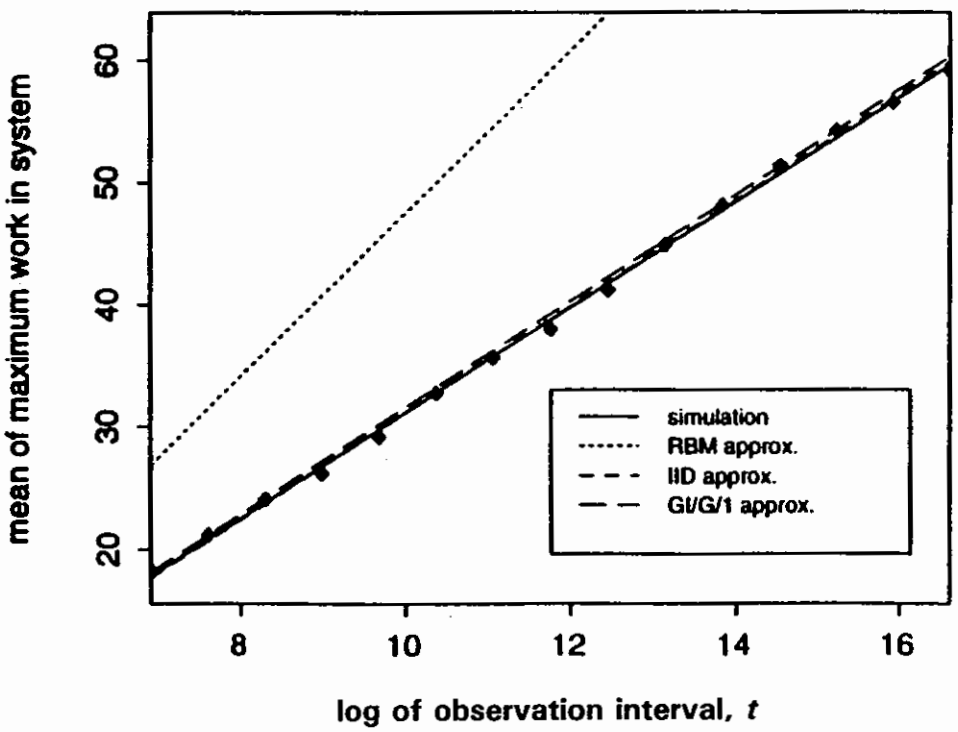

FIGURE 5. $H_2/D/1$, $\rho = 0.7$.

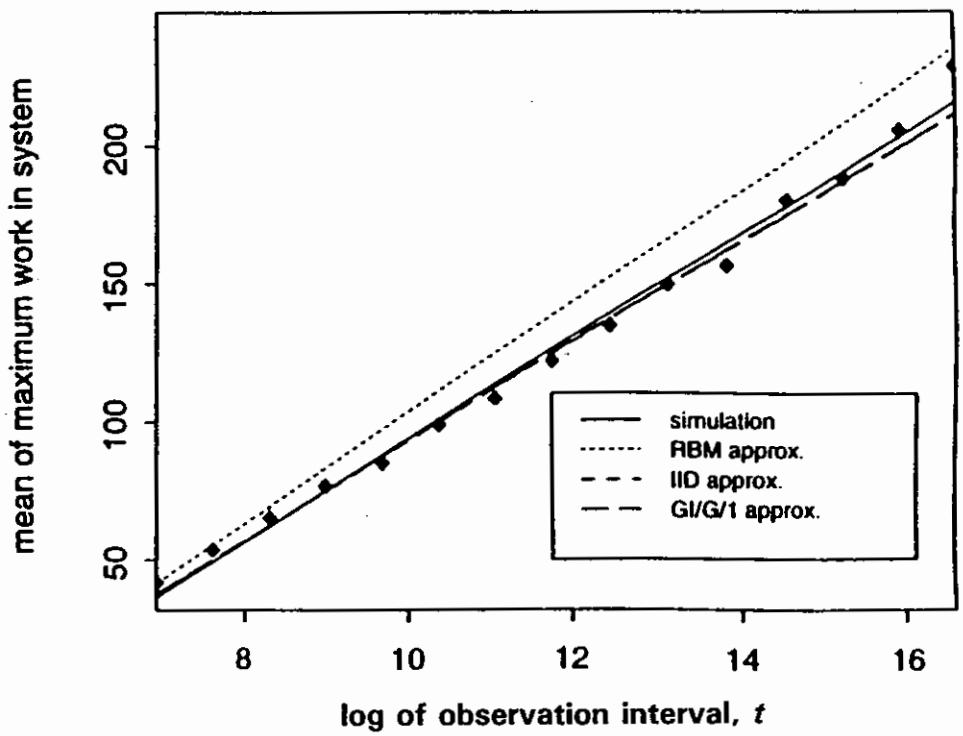

FIGURE 6. $H_2/D/1$, $\rho = 0.9$.

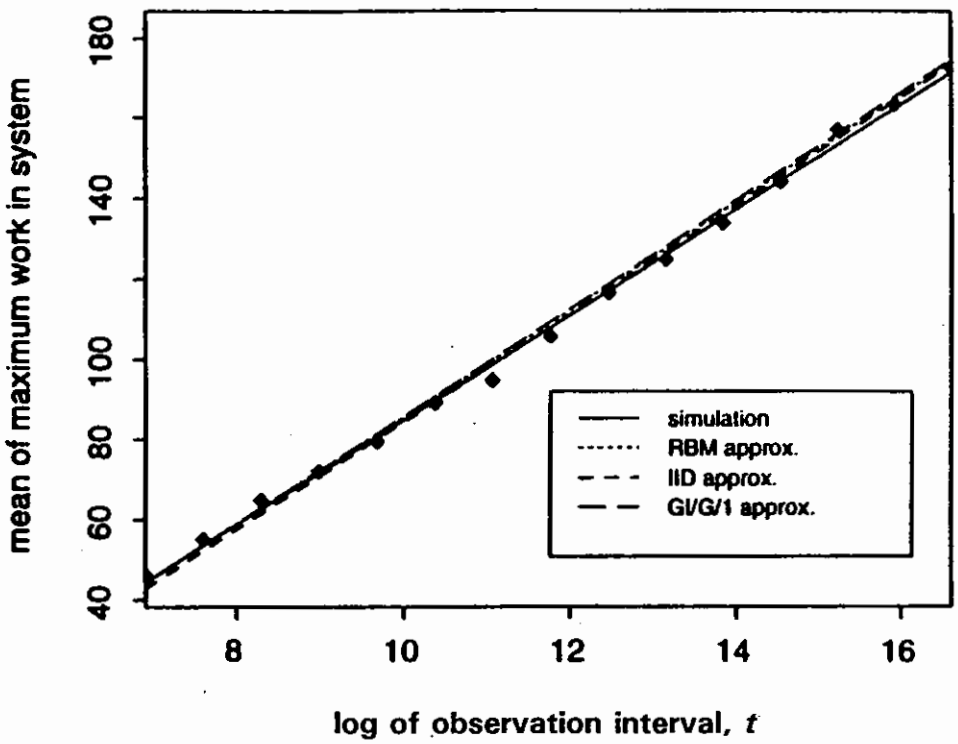

FIGURE 7. $H_2/H_2/1$, $\rho = 0.7$.

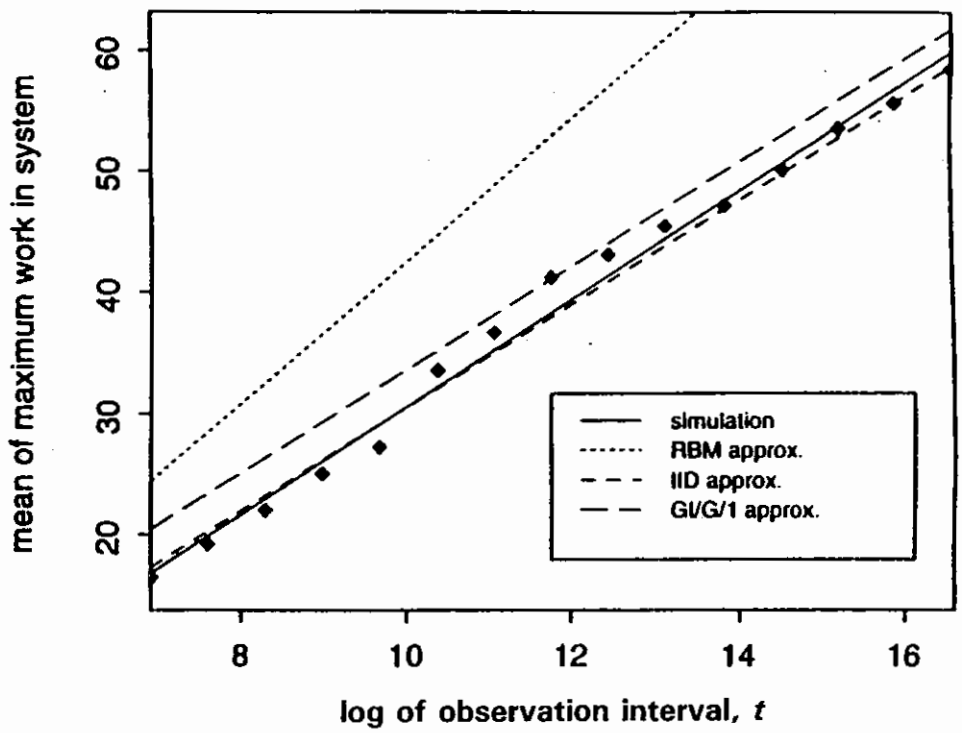

FIGURE 8. MMPP/D/1, $\rho = 0.7$.

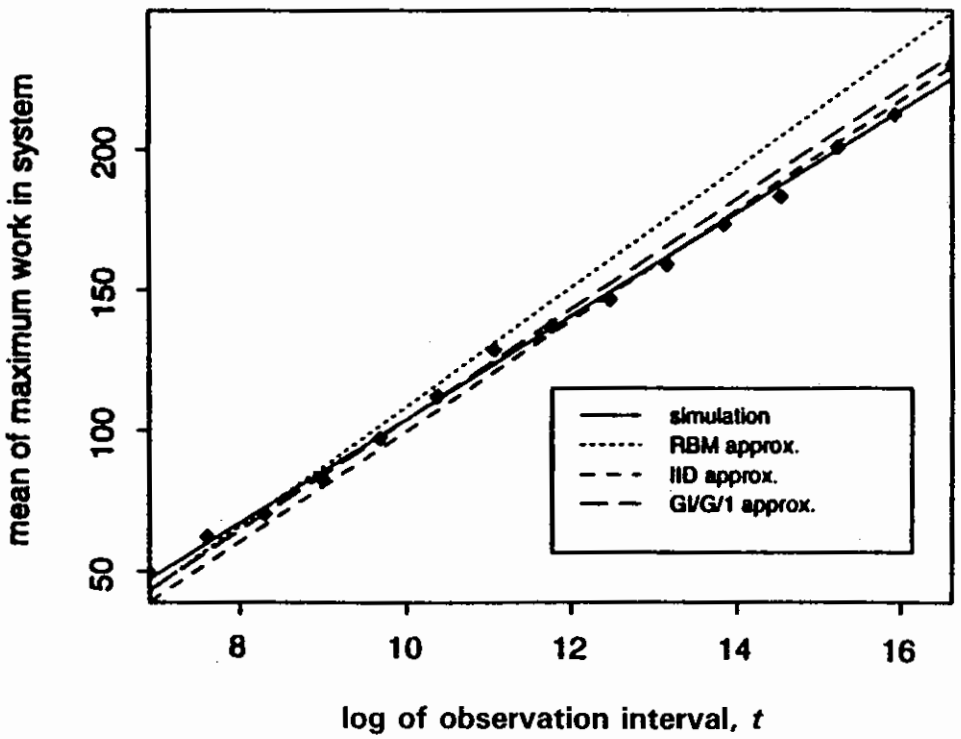

FIGURE 9. MMPP/D/1, $\rho = 0.9$.

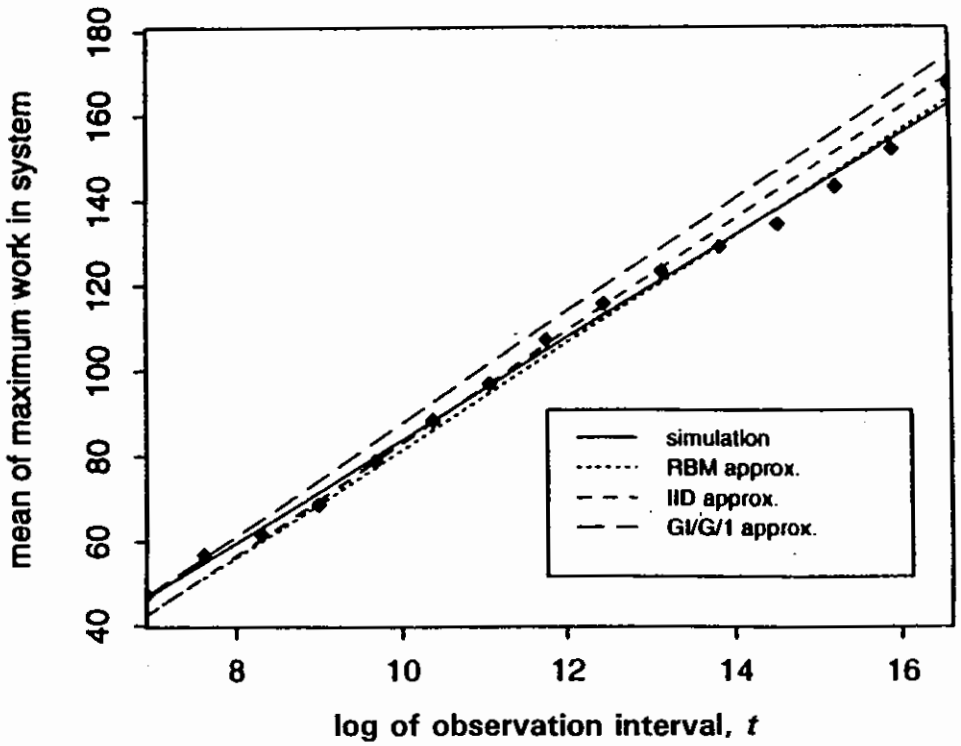

FIGURE 10. MMPP/H_2/1, $\rho = 0.7$.

TABLE 2. A Comparison of Approximations with the Exact Asymptotic Value and Simulation Estimates of the Slope γ in Eq. (1.3) for the Workload Process in Several G/GI/1 Models

Model	Traffic Intensity	Simulation Estimate	$\gamma = 1/\eta$ Exact Asymptotic	RBM Approximation	Mean Steady-State Workload	Asymptotic Constant α
M/M/1	0.7	3.33 ± 0.065	3.33	3.33	2.33	0.7000
	0.9	9.99 ± 0.37	10.00	10.00	9.00	0.9000
M/D/1	0.7	1.43 ± 0.064	1.48	1.67	1.17	0.7990
	0.9	4.62 ± 0.14	4.83	5.00	4.50	0.9333
$M/H_2/1$	0.7	9.68 ± 0.37	10.00	8.33	5.83	0.5727
	0.9	26.0 ± 1.0	26.56	25.00	22.50	0.8452
$H_2/D/1$	0.7	4.31 ± 0.10	4.35	6.67	3.16	0.7269
	0.9	18.49 ± 1.0	17.94	20.00	16.28	0.9074
$H_2/H_2/1$	0.7	13.08 ± 0.38	13.43	13.33	8.36	0.6163
	0.9	39.78 ± 1.8	40.03	40.00	34.83	0.8693
MMPP/D/1	0.7	4.43 ± 0.26	4.26	5.87	2.79	0.6443
	0.9	18.33 ± 0.52	19.63	21.20	17.04	0.8667
$MMPP/H_2/1$	0.7	11.97 ± 0.46	13.12	12.53	7.90	0.5905
	0.9	41.56 ± 1.9	41.44	41.20	35.61	0.8575
Average percent difference			3.0%	10.4%	21.0%	

Eqs. (6.9) and (6.13). Table 4 describes alternative approximations for the y-intercept using c_a^2 instead of c_A^2 and/or the exact value $P(W = 0)$. From Table 4 we see that formula (6.9) with c_a^2 and c_A^2 tend to bound the exact probability $P(W > 0)$ above and below. Overall, we find that all these approximations for the y-intercept are roughly reasonable, without any one clearly dominating the others.

Table 5 compares the approximations for the standard deviation based on Eq. (1.6) with the simulation estimates. We do not display figures corresponding to Figures 3–10 for the standard deviations, because the remaining cases are similar to the M/M/1 case displayed in Figure 2. These approximations also perform well, with the exception that RBM does not predict γ well for D service times and $\rho = 0.7$. Also notice that neither approximation for $\rho = 0.9$ in the $H_2/H_2/1$ case is close; we attribute this largely to variability in the simulation. (This is supported by our experience: A repeat of the simulation with a different set of seeds yielded a sample mean of 56.8 for the standard deviation of maximum work.)

The experiment we have just described shows that the linear relations in Eqs. (1.5) and (1.6) and the good approximations are remarkably accurate for

Table 3. A Comparison of Approximations with Simulation Estimates of the y-intercept in the Linear Relation for the Maximum Workload in Several G/GI/1 Models (the Approximation for log β Appears in Parentheses in the Approximations)

Model	Traffic Intensity	Simulation Estimate	GI/G/1 Approximation	Associated i.i.d. with Eq. (5.9)	RBM	
					Full	With Asymptotic γ
M/M/1	0.7	−7.82 ± 0.79	−7.29 (−2.77)	−7.29 (−2.77)	−7.29 (−2.77)	−7.29
	0.9	−41.9 ± 4.4	−41.3 (−4.71)	−41.3 (−4.71)	−41.3 (−4.71)	−41.3
M/D/1	0.7	−0.92 ± 0.77	−1.84 (−1.82)	−2.02 (−1.94)	−2.49 (−2.07)	−2.21
	0.9	−13.5 ± 1.7	−16.3 (−3.95)	−16.4 (−3.98)	−17.2 (−4.02)	−16.6
M/H_2/1	0.7	−31.1 ± 4.5	−34.9 (−4.06)	−33.0 (−3.88)	−25.9 (−3.68)	−31.0
	0.9	−132.0 ± 12.0	−137.4 (−5.75)	−135.8 (−5.69)	−126.2 (−5.63)	−134.2
H_2/D/1	0.7	−11.9 ± 1.6	−11.9 (−3.32)	−12.4 (−3.42)	−19.2 (−3.46)	−12.5
	0.9	−91.0 ± 13.0	−86.2 (−5.38)	−86.5 (−5.40)	−96.5 (−5.40)	−86.5
H_2/H_2/1	0.7	−45.8 ± 4.6	−48.3 (−4.18)	−49.7 (−4.28)	−47.7 (−4.15)	−48.0
	0.9	−220.0 ± 22.0	−215.0 (−5.95)	−222.0 (−6.13)	−221.0 (−6.10)	−221.0
MMPP/D/1	0.7	−13.8 ± 3.2	−8.97 (−2.69)	−12.1 (−3.41)	−16.2 (−3.33)	−11.7
	0.9	−79.3 ± 6.4	−92.4 (−5.29)	−96.6 (−5.50)	−103.6 (−5.46)	−95.9
MMPP/H_2/1	0.7	−36.1 ± 5.6	−43.7 (−3.91)	−48.3 (−4.26)	−44.0 (−4.09)	−46.1
	0.9	−222.0 ± 23.0	−220.0 (−5.88)	−232.0 (−6.18)	−229.0 (−6.13)	−230.0

TABLE 4. A Comparison of Approximations with the Exact Values of $P(W > 0)$ and with the Simulation Estimates of the y-intercept in the MMPP/G/1 Models

Model	MMPP/D/1		MMPP/H_2/1	
Traffic intensity	$\rho = 0.7$	$\rho = 0.9$	$\rho = 0.7$	$\rho = 0.9$
$P(W > 0)$				
Exact	0.812	0.945	0.775	0.933
Eq. (6.9) and c_a^2	0.760	0.925	0.741	0.916
Eq. (6.9) and c_A^2	0.842	0.958	0.820	0.949
y intercept	−13.81	−79.3	−36.1	−222
Simulation estimate	±3.2	±6.4	±5.6	±23
y intercept				
Exact $P(W = 0)$				
Plus c_a^2 in Eq. (6.13)	−10.87	−101.7	−48.9	−239
Plus c_A^2 in Eq. (6.9)	−8.24	−87.0	−40.8	−208
y intercept				
Eqs. (6.9) and (6.13)				
With c_a^2	−9.82	−95.5	−47.1	−230
With c_A^2	−8.97	−92.4	−43.7	−220

the times we consider. It is natural to ask next how small the times can be and still have these properties. From Figures 1–10 we see that the relations hold reasonably well at the initial time points. To examine this question further, we performed additional experiments with shorter times. We now describe an experiment involving the M/M/1 model with different traffic intensities. We consider 50 independent replications of 100,000 time units, allowing 100,000 time units to warm up to reach steady state before collecting data. (For such shorter times the initial conditions obviously play a bigger role.)

From this experiment, we find that the local slope of the simulation estimates decreases as the time t decreases (i.e., the true curve should be convex). To illustrate, we display the estimates of mean maximum workload for the cases $\rho = 0.7$ and $\rho = 0.9$ in Figures 11 and 12. For each time point we include estimates of the 95% confidence intervals. From such figures it is tempting to estimate when the asymptotics do in fact begin to take effect. The experiment shows that the change is continuous, but nevertheless we try to estimate the knee of the curve. Our rough estimates of the knee of the curve appear in Figures 11 and 12 plus Table 6. For the case $\rho = 0.7$ the slope of the regression line through the first 31 points (below the estimated knee) is 1.85, while the slope of the

Table 5. A Comparison of the Mean of the 15 Sample Standard Deviations with the Approximations Based on Eq. (1.6)

			$1.28/\eta$	
Model	Traffic Intensity	Simulation Estimate	Exact Asymptotic	RBM
M/M/1	0.7	3.72	4.26	4.26
	0.9	11.40	12.80	12.80
M/D/1	0.7	1.88	1.89	2.14
	0.9	6.23	6.18	6.40
$M/H_2/1$	0.7	12.31	12.80	10.66
	0.9	31.88	34.00	32.00
$H_2/D/1$	0.7	4.73	5.57	8.54
	0.9	22.36	22.96	25.60
$H_2/H_2/1$	0.7	18.37	17.19	17.06
	0.9	38.48	51.24	51.20
MMPP/D/1	0.7	5.50	5.45	7.51
	0.9	22.10	25.13	27.14
$MMPP/H_2/1$	0.7	16.82	16.80	16.04
	0.9	53.44	53.04	52.74

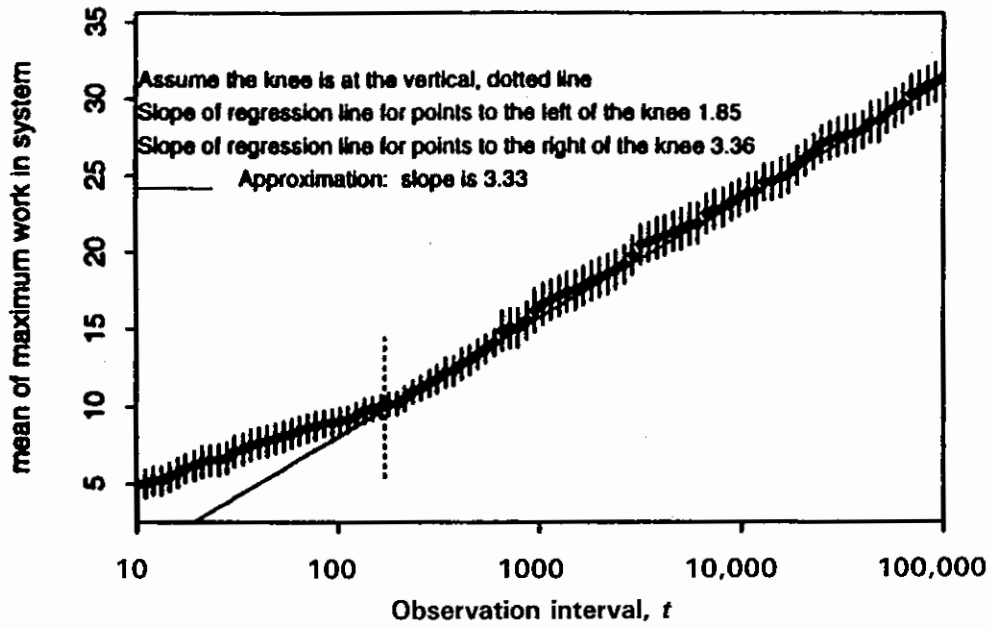

Figure 11. Mean of maximum work in system realized over 50 sample paths. M/M/1, $\rho = 0.7$.

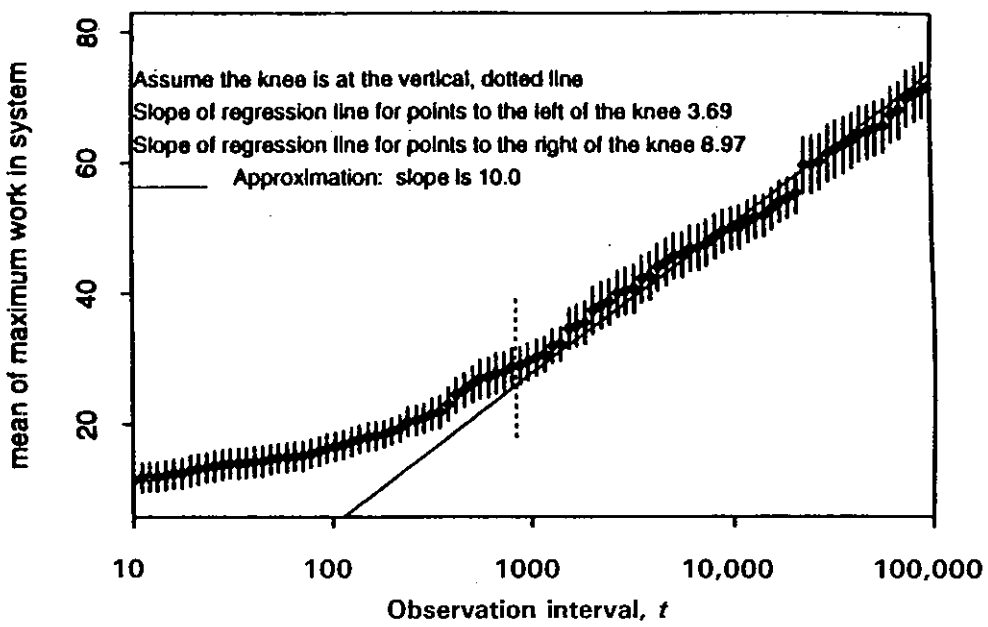

FIGURE 12. Mean of maximum work in system realized over 50 sample paths. M/M/1, $\rho = 0.9$.

regression line through the last 69 points (above the estimated knee) is 3.36, where $\eta^{-1} = 3.33$. For the case $\rho = 0.9$, the slope of the regression line through the first 48 points (below the estimated knee) is 3.69, while the slope of the regression line through the last 52 points (above the estimated knee) is 8.97,

TABLE 6. Estimated Location of the Knee of the Maximum Mean Workload Curves for the M/M/1 and $H_2/H_2/1$ Models as a Function of the Traffic Intensity ρ

		Location of Knee				
Model	Traffic Intensity	Number of Points from Left	Time	Knee $\times (1-\rho)$	Knee $\times (1-\rho)^2$	Knee $\times \dfrac{(1-\rho)^2}{4(c_a^2 + c_s^2)}$
M/M/1	0.5	14	33.5	16.7	8.4	1.05
	0.7	31	163	48.9	14.7	1.84
	0.9	48	793	79.3	7.9	0.99
$H_2/H_2/1$	0.7	43	498	149	44.8	1.40
	0.9	60	2420	242	24.2	0.76

where $\eta^{-1} = 10.00$. Hence, we conclude that indeed the estimated knee is approximately where the extreme value limits begin to take effect.

The RBM approximation in Eq. (4.11) suggests that the knee should be proportional to $b/2(1-\rho)^2a^2$, because that is how the knee for RBM would be transformed to the queueing processes. This estimate is supported by the fact that, for the estimates, the knee $\times (1-\rho)^2$ is approximately constant, whereas knee $\times (1-\rho)$ is not. *As a tentative general rough approximation for the time where the approximation begins to take effect, drawing on Eq. (4.3), we suggest the associated* RBM *approximation*

$$\text{knee of curve} \approx \frac{4b}{a^2(1-\rho)^2}, \tag{7.2}$$

which is $8/(1-\rho)^2$ *for* M/M/1. This rough approximation is supported by Table 6.

Qualitatively, this approximation formula for the knee implies that the knee is increasing in both ρ and the model variability (through the parameter b). The first property is evident from Figures 11 and 12. The second property is evident from similar figures for the other models we have considered, such as $H_2/H_2/1$. The case of $H_2/H_2/1$ with $\rho = 0.7$ and $\rho = 0.9$ is shown in Figures 13 and 14.

We conclude this empirical section by considering a case for which extreme-value limits support a *different* approximation than Eq. (1.3). In particular, we consider an M/G/1 queue with a Pareto service-time density

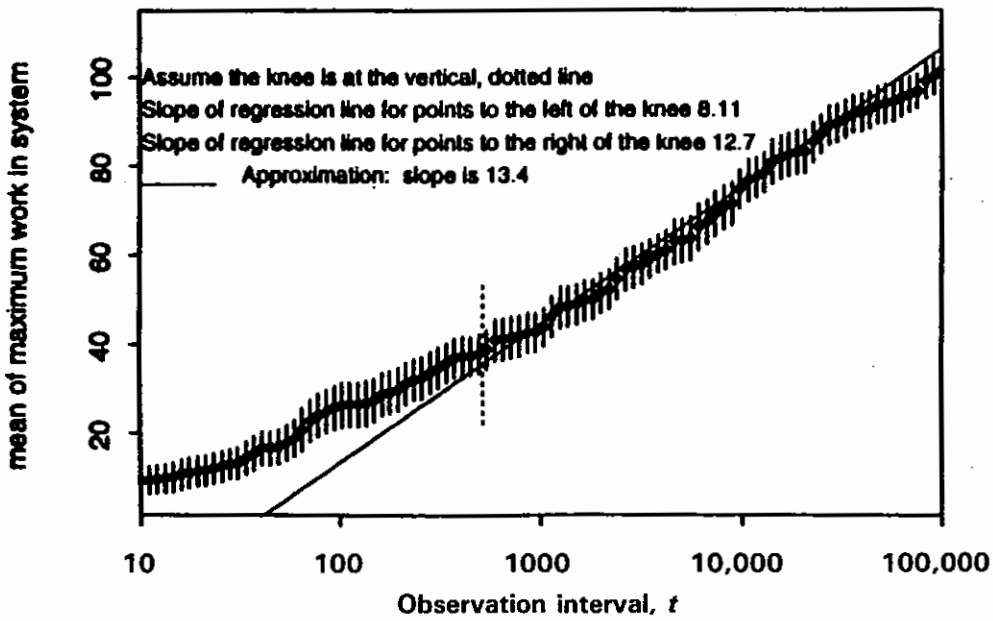

FIGURE 13. Mean of maximum work in system realized over 50 sample paths. $H_2/H_2/1$, $\rho = 0.7$.

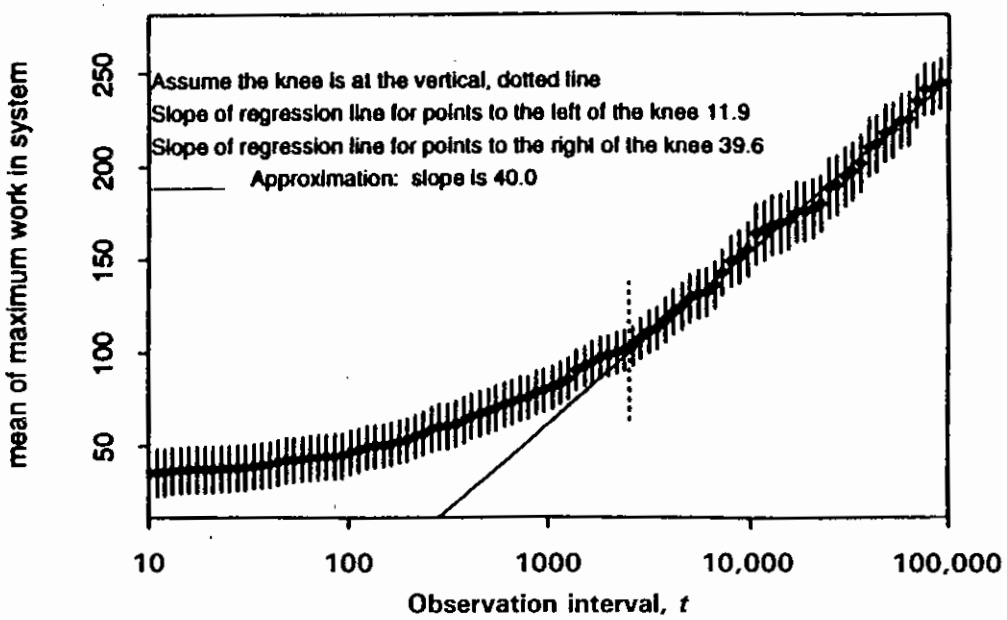

FIGURE 14. Mean of maximum work in system realized over 50 sample paths. $H_2/H_2/1$, $\rho = 0.9$.

$$f_r(x) = r\left(\frac{r-1}{r}\right)^r x^{-(r+1)}, \qquad x \geq (r-1)/r, \tag{7.3}$$

which has a mean of 1 and $c_s^2 = 1/r(r-2)$ (see Johnson and Kotz [26, Ch. 22] and Abate et al. [3, Section 2]).

Paralleling Figure 1, Figure 15 plots the sample means of the maximum workloads for 20 independent replications in the case $r = 3.5$ ($c_s^2 = 0.19$) and $\rho = 0.9$. In contrast to the previous examples, Figure 15 clearly shows that the linear relation in Eq. (1.5) does *not* hold for this example: in this case, the mean of the maximum workload is *not* linear in $\log t$ for large t.

However, consistent with the discussion in Section 4 following Eq. (4.6), the linear relation may hold for a range of times neither too small nor too large. For example, there is a pretty good linear fit to the first 9 points in Figure 15. A regression analysis yields an estimated slope of 6.2, while the RBM approximation based on Eq. (4.11) indicates a slope of 6.0.

For this example, the extreme-value theory [29] indicates that the maximum workload should in fact be linear in $t^{1/2.5}$ instead of $\log t$. Hence, in Figure 16 we replot Figure 15 in this scale. Again we see that there is not linearity for all times, but there seems to be a linear relation for the last 6 points. Consistent with experience with steady-state tail probabilities [3], a comparison of Figures 1 and 16 indicates that the extreme-value asymptotics does not take effect as quickly with long-tail service-time distributions such as the Pareto distribution as it does for the "standard" service-time distributions in the domain of approx-

2. Abate, J., Choudhury, G.L., & Whitt, W. (1994). Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models* 10: 99–143.
3. Abate, J., Choudhury, G.L., & Whitt, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems* 16: 311–338.
4. Abate, J., Choudhury, G.L., & Whitt, W. (to appear). Exponential approximations for tail probabilities in queues, I: Waiting times. *Operations Research.*
5. Abate, J., Choudhury, G.L., & Whitt, W. (to appear). Exponential approximations for tail probabilities in queues, II: Sojourn time and workload. *Operations Research.*
6. Abate, J. & Whitt, W. (1987). Transient behavior of regulated Brownian motion, I: Starting at the origin. *Advances in Applied Probability* 19: 560–598.
7. Abate, J. & Whitt, W. (1988). Transient behavior of the M/M/1 queue via Laplace transforms. *Advances in Applied Probability* 20: 145–178.
8. Asmussen, S. (1992). Queueing simulations in heavy traffic. *Mathematics of Operations Research* 17: 84–111.
9. Asmussen, S. & Perry, D. (1992). On cycle maxima, first passage problems and extreme value theory of queues. *Stochastic Models* 8: 421–458.
10. Bailey, N.T.J. (1957). Some further results in the non-equilibrium theory of a simple queue. *Journal of the Royal Statistical Society B* 19: 326–333.
11. Berger, A.W. & Whitt, W. (1992). The Brownian approximation for rate-control throttles and the G/G/1/C queue. *Discrete Event Dynamic Systems* 2: 7–60.
12. Berger, A.W. & Whitt, W. (1994). Asymptotics for open-loop window flow control. *Journal of Applied Mathematics and Stochastic Analysis* 7: 337–356.
13. Berger, A.W. & Whitt, W. (to appear). Comparison of the sliding window and the leaky bucket. *Queueing Systems.*
14. Berman, S. (1962). Limiting distributions of the maximum term in sequences of dependent random variables. *Annals of Mathematical Statistics* 33: 894–908.
15. Billingsley, P. (1968). *Convergence of probability measures.* New York: Wiley.
16. Castillo, E. (1988). *Extreme value theory in engineering.* Boston: Academic Press.
17. Choudhury, G.L. & Lucantoni, D.M. (to appear). Numerical computation of the moments of a probability distribution from its transforms. *Operations Research.*
18. Choudhury, G.L., Lucantoni, D.M., & Whitt, W. (to appear). Squeezing the most out of ATM. *IEEE Transactions on Communications.*
19. Choudhury, G.L. & Whitt, W. (1994). Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 queue. *Stochastic Models* 10: 453–498.
20. Cohen, J.W. (1968). Extreme value distributions for the M/G/1 and GI/M/1 queueing systems. *Annales de l'Institut Henri Poincaré Sect. B* 4: 83–98.
21. Darling, D.A. & Siegert, A.J.F. (1953). The first passage problem for a continuous Markov process. *Annals of Mathematical Statistics* 24: 624–639.
22. Glynn, P.W. & Whitt, W. (to appear). Heavy-traffic extreme-value limits for queues. *Operations Research Letters.*
23. Halfin, S. (1985). Delays in queues, properties and approximations. In M. Akiyama (ed.), *Teletraffic issues in an advanced information society,* ITC 11. Amsterdam: Elsevier, pp. 47–52.
24. Harrison, J.M. & Nguyen, V. (1993). Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems* 13: 5–40.
25. Iglehart, D.L. (1972). Extreme values in the GI/G/1 queue. *Annals of Mathematical Statistics* 43: 627–635.
26. Johnson, N.L. & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions-1.* New York: Wiley.
27. Kemeny, J.G. & Snell, J.L. (1960) *Finite Markov chains.* New York: Van Nostrand.
28. Kraemer, W. & Langenbach-Belz, M. (1976). Approximate formulae for the delay in the queueing system GI/G/1. *Proceedings of the Eighth International Teletraffic Congress,* Melbourne, 235-1/8.

29. Leadbetter, M.R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes.* New York: Springer-Verlag.
30. McCormick, W.P. & Park, Y.S. (1992). Approximating the distribution of the maximum queue length for M/M/s queues. In U.N. Bhat & I.V. Basawa (eds.), *Queueing and related models.* Oxford: Oxford University Press, pp. 240–261.
31. Neuts, M.F. (1989). *Structured stochastic matrices of M/G/1 type and their applications.* New York: Marcel Dekker.
32. Pakes, A.G. (1975). On the tails of waiting-time distributions. *Journal of Applied Probability* 12: 555–564.
33. Sadowsky, J.S. (1995). The probability of large queue lengths and waiting times in a heterogeneous multiserver queue, Pt. II: Positive recurrence and logarithmic limits. *Advances in Applied Probability* 27: 567–583.
34. Sadowsky, J.S. & Szpankowski, W. (1992). Maximum queue length and waiting time revisited: G/G/c queue. *Probability in the Engineering and Informational Sciences* 6: 157–170.
35. Sadowsky, J.S. & Szpankowski, W. (1995). The probability of large queue lengths and waiting times in a heterogeneous multiserver queue, Pt. I: Tight limits. *Advances in Applied Probability* 27: 532–566.
36. Serfozo, R.F. (1988). Extreme values of birth and death processes and queues. *Stochastic Processes and Their Applications* 27: 291–306.
37. Serfozo, R.F. (1988). Extreme values of queue lengths in M/G/1 and GI/M/1 systems. *Mathematics of Operations Research* 13: 349–357.
38. Whitt, W. (1989). Planning queueing simulations. *Management Science* 35: 1341–1366.
39. Whitt, W. (1992). Asymptotic formulas for Markov processes with applications to simulations. *Operations Research* 40: 279–291.
40. Whitt, W. (1993). Approximations for the GI/G/m queue. *Production and Operations Management* 2: 114–161.