LARGE DEVIATIONS FOR THE INFINITE SERVER QUEUE IN HEAVY TRAFFIC

PETER W. GLYNN*

Abstract. In this paper, we establish large deviations approximations to tail probabilities of the queue-length r.v. in an infinite-server queue in heavy traffic. These large deviations approximations complement the existing Gaussian approximations developed for such systems using weak convergence theory on function spaces. We also describe a simulation-based algorithm for numerically computing such tail probabilities that takes advantage of the large deviations theory developed here.

Key words. infinite server queue, large deviations, simulation

1. Introduction. This paper is concerned with developing large deviations approximations for tail probabilities of infinite server queues in heavy traffic. By heavy traffic, we refer to systems in which the offered load is high, so that the system typically contains many customers.

The theory developed here is basically a many-server analog to the Cramer-Lundberg type approximations that are commonly used to study the single-server queue. This paper has two parts. In Section 2, we develop a large deviations approximation for the tail probabilities associated with the number of customers in the system at time t when the system is in heavy traffic. We also provide a large deviations argument to offer asymptotic justification for the approximation. Section 3 is concerned with simulation-based algorithms for numerically computing the tail probabilities that take advantage of the large deviations ideas presented in Section 2.

An expanded version of this paper will appear elsewhere.

2. Large deviations for $Q(t)$. We start by giving a precise description of the $GI/G/\infty$ queue. Suppose that $(A_k : k \geq 0)$ is a non-decreasing sequence in which A_k corresponds to the arrival epoch of the k'th customer. If $Q(0) = 0$ and V_j is the "time-in-system" of the j'th customer, then the number of customers $Q(t)$ in the system at time t is given by

$$Q(t) = \sum_{k=1}^{\infty} I(A_k \leq t < A_k + V_k).$$

Let $N(t) = \max\{n \geq 0 : A_n \leq t\}$ be the number of customers to arrive to the system in $[0, t]$. Then, $Q(t)$ can be re-expressed in terms of $N(\cdot)$ as

$$Q(t) = \sum_{k=1}^{N(t)} I(A_k + V_k > t).$$

* Dept. of Operations Research, Stanford University, Stanford, CA 94305-4022.
This research was supported by the National Science Foundation under grant DDM-9101580 and the Army Research Office under Contract No. DAAL03-91-G-0319.

We will be using the Gärtner-Ellis theorem to study the large deviations behavior of the r.v. $Q(t)$. In order to apply this result, the moment generating function of $Q(t)$ must be computed. This calculation is particularly simple when the sequence $V = (V_n : n \geq 1)$ is i.i.d. and independent of $N = (N(t) : t \geq 0)$; we therefore impose this condition throughout the paper. Let $F(x) = P(V \leq x)$ and $\bar{F}(x) = 1 - F(x)$.

PROPOSITION 2.1. *Suppose that* $\varphi(\theta, t) = E \exp(\theta Q(t))$. *Then,*

$$\varphi(\theta,t) = E\exp\Big(\int_{[0,t]} \log(e^{\theta}\,\bar{F}(t-x) + F(t-x))\; N(dx)\Big). \tag{2.1}$$

Proof. Note that

$$\begin{aligned}
\varphi(\theta,t) &= E\{\,E[\exp(\theta Q(t)) \mid N]\,\} \\
&= E\{\,E[\exp(\theta \sum_{k=1}^{N(t)} I(A_k + V_k > t))\mid N]\,\} \\
&= E\prod_{k=1}^{N(t)} E[\exp(\theta\, I(V_k > t - A_k))\mid N] \\
&= E\prod_{k=1}^{N(t)} (e^{\theta}\bar{F}(t-A_k) + F(t-A_k)) \\
&= E\exp(\sum_{k=1}^{N(t)} \log(e^{\theta}\bar{F}(t-A_k) + F(t-A_k))) \\
&= E\exp\Big(\int_{[0,t]} \log(e^{\theta}\bar{F}(t-x) + F(t-x))\, N(dx)\Big).
\end{aligned}$$

□

Large deviations theory is intended here to provide refined approximations, relative to the central limit theory already developed, for the tail probabilities of the r.v. $Q(t)$. By "central limit theory", we refer here to the idea that if the arrival rate is high, then various limit theorems make rigorous the approximation

$$Q(t) \overset{\mathcal{D}}{\approx} EQ(t) + \sqrt{\mathrm{var}Q(t)}\, N(0,1), \tag{2.2}$$

where $\overset{\mathcal{D}}{\approx}$ denotes approximate equality in distribution, and $N(0,1)$ is a standard normal r.v.; see Iglehart (1965), Borovkov (1967), (1984), Newell (1973), Whitt (1982), and Glynn and Whitt (1991) for details. In particular, the approximation to $P(Q(t) > x)$ suggested by (2.2) is typically good whenever the arrival rate is high, and $|x - EQ(t)|/\sqrt{\mathrm{var}Q(t)}$ is of moderate magnitude. (Note that there is no requirement here that t be large, in order that (2.2) yield a good approximation.) A reasonable heuristic for judging when the arrival rate is high is to look for situations in which the standard deviation $\sqrt{\mathrm{var}Q(t)}$ is small relative to $EQ(t)$.

Large deviations is intended to provide improved approximations to $P(Q(t) > x)$ when the arrival rate is high, and $|x - EQ(t)|/\sqrt{\mathrm{var}Q(t)}$ is large. Let $\psi(\theta, t) = \log \varphi(\theta, t)$ and suppose that θ^* is the root (assumed to exist uniquely) of $\psi'(\theta^*, t) = x$. Then, large deviations suggests the rough approximation

$$P(Q(t) > x) \approx \exp(-\theta^* x + \psi(\theta^*, t)). \tag{2.3}$$

Our goal is to now describe several asymptotic regimes in which (2.3) is valid, and describe the precise sense in which the approximation holds.

To accomplish this, we consider a sequence of systems in which the arrival rate is sent to infinity. Let $N_n = (N_n(t) : t \geq 0)$ be the arrival process to the n'th system, so that $N_n(t)$ represents the number of customers to arrive to the n'th system in $[0, t]$. We leave the service time sequence $V = (V_n : n \geq 1)$ unchanged as a function of n, and let $Q_n(t)$ be the corresponding number of customers in the n'th system at time t, so that

$$Q_n(t) = \sum_{k=1}^{N_n(t)} I(A_{k,n} + V_k > t),$$

where $A_{k,n} = \min\{t \geq 0 : N_n(t) \geq k\}$.

There are two fundamentally different ways of modeling the notion that the arrival rate to the system is high. Most of the "heavy traffic" literature assumes the presence of a single "fast" source. The easiest mathematical means of studying this situation is to fix a given arrival process N, and to speed up the arrival rate via the sequence of processes

$$N_n(t) = N(nt).$$

Here, $A_{k,n} = A_k/n$ and the arrival rate associated with $N_n(\cdot)$ is roughly n times that of $N(\cdot)$. We will make the following assumption about the joint cumulant generating function of the increments of N.

A1. *There exists a finite-valued function* ψ_N *such that for* $0 = t_0 < t_1 < \ldots < t_m = t$ *and* $(\theta_1, \ldots, \theta_m) \in \Re^m$,

$$\frac{1}{n}\log E\exp(\sum_{i=1}^{m} \theta_i[N(nt_i) - N(nt_{i-1})]) \to \sum_{i=1}^{m} \psi_N(\theta_i)(t_i - t_{i-1})$$

as $n \to \infty$.

This assumption is satisfied by many different arrival processes; see Dembo and Zajic (1993). To offer some insight into this assumption, we describe the function ψ_N in a couple of important applications settings:

Example 2.2. Suppose that the arrival process is renewal, so that A_k can be represented as $A_k = U_1 + \ldots + U_k$, where $(U_k : k \geq 1)$ is i.i.d. Then, under suitable regularity conditions on the U_k's (see Glynn and Whitt (1994)),

$$\psi_N(\theta) = -\kappa^{-1}(-\theta),$$

where $\kappa(\theta) = \log(\, E \exp(\theta U_1)\,)$.

Example 2.3. Suppose now that the arrival process is a Markov-modulated Poisson process, so that there exists an S-valued continuous-time Markov chain $X = (X(t) : t \geq 0)$ and a function $f : S \to (0, \infty)$ such that the arrival rate of the Poisson process at time t is $f(X(t))$. Assume that S is finite and X is irreducible. Then, if B is the generator of X, set

$$B(\theta) = B + D(\theta)$$

where $D(\theta) = \text{diag}((e^\theta - 1)f(x) : x \in S)$. Here, $\psi_N(\theta)$ is the eigenvalue of $B(\theta)$ having maximum real part (which turns out necessarily to be real).

In any case, to study the large deviations behavior of $Q_n(t)$ via the Gärtner-Ellis theorem, we must obtain the limit behavior of the cumulant generating function of $Q_n(t)$ defined by $\psi_n(\theta, t) = \log E \exp(\theta Q_n(t))$.

THEOREM 2.4. *If A1 holds, then*

$$\frac{1}{n}\psi_n(\theta, t) \to \int_0^t \psi_N(\log(e^\theta \bar{F}(x) + F(x)))\, dx \tag{2.4}$$

as $n \to \infty$.

Proof. Suppose $\theta \geq 0$. Then, $h(x) = \log(e^\theta \bar{F}(t - x) + F(t - x))$ is non-decreasing in x. We therefore obtain the bounds

$$\begin{aligned}
&\frac{1}{n}\log E \exp\Big(\sum_{k=0}^{m-1} h(\frac{kt}{m})\,[\, N(\frac{n(k+1)t}{m}) - N(\frac{nkt}{m})\,]\,\Big) \\
&\leq \frac{1}{n}\log E \exp\Big(\int_{[0,t]} \log(e^\theta \bar{F}(t - x) + F(t - x))\, N_n(dx)\Big) \\
&\leq \frac{1}{n}\log E \exp\Big(\sum_{k=0}^{m-1} h(\frac{(k+1)t}{m})\,[\, N(\frac{n(k+1)t}{m}) - N(\frac{nkt}{m})\,]\,\Big).
\end{aligned}$$

Letting $n \to \infty$ and using Proposition 2.1 as well as A1, we get

$$\begin{aligned}
\frac{t}{m}\sum_{k=0}^{m-1} \psi_N(h(\frac{kt}{m})) &\leq \liminf_{n\to\infty} \frac{1}{n}\psi_n(\theta, t) \\
&\leq \limsup_{n\to\infty} \frac{1}{n}\psi_n(\theta, t) \leq \frac{t}{m}\sum_{k=0}^{m-1} \psi_N(h(\frac{(k+1)t}{m})).
\end{aligned} \tag{2.5}$$

Now, note that by setting $n = 1$ in A1, it is clear that ψ_N, being the limit of convex functions, must necessarily be convex and hence continuous. It follows that the extreme members of the inequality (2.5) are Riemann approximations to the integral of the continuous function $\psi_N \circ h$. By letting $m \to \infty$, we therefore conclude that

$$\frac{1}{n}\psi_n(\theta, t) \to \int_0^t \psi_N(h(x))dx,$$

as desired. For $\theta < 0$, a similar argument works. □

Let

$$\psi_{Q(t)}(\theta) = \int_0^t \psi_N(\,\log(e^\theta \bar{F}(x) + F(x)))\, dx.$$

Suppose that $\psi_{Q(t)}(\cdot)$ is differentiable and that there exists a unique θ^* such that $\psi'_{Q(t)}(\theta^*) = x$. Then, the Gärtner-Ellis theorem implies that

$$\frac{1}{n}\log P(Q_n(t) > xn) \to -\theta^* x + \psi_{Q(t)}(\theta^*)$$

as $n \to \infty$; see, for example, p.14–15 of Bucklew (1990). On the other hand, if $x_n = xn$, (2.4) establishes that

$$\frac{1}{n}\log(\, \exp(-\theta^* x_n + \psi_n(\theta^*, t))\,) \to -\theta^* x + \psi_{Q(t)}(\theta^*).$$

Furthermore, since the ψ_n's are convex, we may differentiate through (2.4) (see Rockafellar (1970)), yielding

$$\frac{1}{n}\psi'_n(\theta, t) \to \psi'_{Q(t)}(\theta)$$

as $n \to \infty$. Since $\psi'_n(\cdot, t)$ is non-decreasing, we can conclude that $\theta_n^* \to \theta^*$ as $n \to \infty$, where θ_n^* is the root of $\psi'_n(\theta_n^*, t) = x_n$. It follows that

$$\frac{1}{x_n}\log[\, P(Q_n(t) > x_n)/\exp(-\theta_n^* x_n + \psi_n(\theta_n^*, t))\,] \to 0 \tag{2.6}$$

as $n \to \infty$. We view (2.6) as a rigorous statement of the mathematical sense in which the right-hand side of (2.3) is a valid approximation to $P(Q(t) > x)$ (when $x/EQ(t)$ is of moderate magnitude, and not close to unity).

A second setting in which (2.3) is valid is when the high arrival rate is due to the presence of many sources, each of moderate size. Mathematically, this situation can be captured by fixing a given arrival process N, and letting $N(1, \cdot), N(2, \cdot), \ldots$ be i.i.d. copies of N. Here, we set

$$N_n(t) = \sum_{i=1}^n N(i, t)$$

for $t \geq 0$. Since

$$Q_n(t) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{n} \int_{[0,t]} I(V_{ij} > t - x)\, N(i, dx)$$

($\stackrel{\mathcal{D}}{=}$ denotes "equality in distribution"), where $(V_{ij} : i \geq 1, j \geq 1)$ is a family of i.i.d. copies of V_1, it is evident that $Q_n(t)$ is a sum of n i.i.d. r.v.'s. To obtain a large deviations result in this setting is purely a matter of applying the existing theory for i.i.d. r.v.'s. This large deviations result implies that the approximation (2.3) can be reasonable when many sources are present (so that $EQ(t)/\sqrt{\text{var}Q(t)}$ is large) and the tail probability $P(Q(t) > x)$ is such that $x/EQ(t)$ is of moderate magnitude and not close to unity; the argument is similar to that leading to (2.6).

As one might expect, these results can easily be extended to large deviations approximations for tail probabilities of the steady-state r.v. $Q(\infty)$. To apply (2.3), note that

$$\varphi(\theta, +\infty) = E \exp\Big(\int_{[0,\infty]} \log(e^{\theta} \bar{F}(x) + F(x))\, \tilde{N}(dx) \Big),$$

where $\tilde{N} = (\tilde{N}(t) : t \geq 0)$ is a time-stationary version of the time-reversal of the arrival process N; details will be supplied elsewhere.

3. Simulation implications. A glance at (2.6) suggests that the right-hand side of (2.3) is typically, at best, only a crude approximation to $P(Q(t) > x)$. In particular, we note that the approximation can be multiplied by any constant factor, without any impact on the logarithmic asymptotic (2.6). Hence, the right-hand side of (2.3) should perhaps only be viewed as an order-of-magnitude guess as to the size of $P(Q(t) > x)$.

Given this state of affairs, it seems reasonable to consider efficient numerical algorithms for computing the tail probability $P(Q(t) > x)$. We shall describe here some "importance sampling" ideas, based on the large deviations theory already developed, for efficiently simulating this tail probability.

The theory of rare event simulation suggests that the tail probability $P(Q(t) > x)$ can be efficiently simulated under the change-of-measure

$$P^*(d\omega) = \exp(\theta^* x - \psi(\theta^*, t))\, P(d\omega), \tag{3.1}$$

where θ^* is as defined in (2.3). Furthermore, the level of variance reduction incurred by using (3.1) increases under the same asymptotic regimes as those validating the approximation (2.3). Of course, the computational efficiency of such a simulation algorithm depends not only on the variance but also on the computational ease associated with producing simulation runs under the modified probability measure P^*. Generating variates under P^* is, in this setting, quite challenging as it typically fails to induce any Markovian structure on the dynamical behavior of the simulated process.

However, there exists a modification to P^*, call it $\tilde{P}$, which enjoys the same asymptotic variance reducing properties as does P^*, and which is easily simulatable. The probability measure $\tilde{P}$ is easiest to describe in the setting in which the process N is a non-delayed renewal process, and we shall therefore specialize to this important subclass of problems. Let $U_n = A_n - A_{n-1}$, $\kappa(\theta) = \log E \exp(\theta U_i)$ (assumed to converge for all θ), and put $\psi(\theta) = -\kappa^{-1}(-\theta)$. The assumption that $\kappa(\theta)$ is finite for all θ is a technical restriction that we impose merely to easily describe the algorithm and result, and should in no way be viewed as necessary.

The following simulation algorithm implicitly describes $\tilde{P}$. The quantities Q and L defined in the algorithm correspond to the r.v.'s $Q(t)$ and $dP/d\tilde{P}$, as generated under $\tilde{P}$.

Algorithm 3.1.

1. $A \leftarrow 0$, $Q \leftarrow 0$, $L \leftarrow 1$.
2. Generate U from the distribution
   $\exp(-\psi(\log(e^{\theta^*}\bar{F}(t-A)+F(t-A)))x+\log(e^{\theta^*}\bar{F}(t-A)+F(t-A)))$
   $P(U_i \in dx)$.
3. $L \leftarrow L \exp(\psi(\log(e^{\theta^*}\bar{F}(t-A)+F(t-A)))U)\,(e^{\theta^*}\bar{F}(t-A)$
   $+F(t-A))^{-1}$.
4. $A \leftarrow A+U$.
5. If $A > t$, return Q, L.
6. Generate the $0-1$ r.v. I from Bernoulli $(e^{\theta^*}\bar{F}(t-A)\,(e^{\theta^*}\bar{F}(t-A)$
   $+F(t-A))^{-1})$.
7. $L \leftarrow L\, e^{-\theta^* I}\,(e^{\theta^*}\bar{F}(t-A)+F(t-A))$.
8. $Q \leftarrow Q+I$.
9. Go to 2.

This algorithm is intended to work efficiently when the system is in "heavy traffic", as a result of a single "fast" renewal arrival source. Full asymptotic justification for this algorithm will appear elsewhere; the algorithm turns out to be asymptotically "optimal" in a certain sense.

We note that when "heavy traffic" is achieved as a result of the presence of many i.i.d. arrival sources, rare event simulation algorithms for tail probabilities of sums of i.i.d. r.v.'s comes into play; for details, see, for example, Bucklew (1990). This theory suggests that each of the sources should be simulated independently. The change-of-measure for the individual sources takes the same form as that specified above in the "fast source" setting.

REFERENCES

[1] Borovkov, A. A. (1967) On limit laws for service processes in multi-channel systems. *Siberian Math. J.* **8**, 746–763.

[2] Borovkov, A. A. (1984) *Asymptotic Methods in Queueing Theory.* Wiley, New York.

[3] Bucklew, J. A. (1990) *Large Deviation Techniques in Decision, Simulation, and Estimation.* J. Wiley and Sons, New York.

[4] Dembo, A. and Zajic, T. (1993) Large deviations: from empirical mean and measure to partial sums process. Preprint.

[5] Glynn, P. W. and Whitt, W. (1991) A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Prob.* **23**, 188–209.

[6] Glynn, P. W. and Whitt, W. (1994) Large deviation behavior of counting processes and their inverses. *Queueing Systems* **17**, 107–128.

[7] Iglehart, D. L. (1965) Limit diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* **2**, 429–441.

[8] Newell, G. F. (1973) *Approximate Stochastic Behavior of n-Server Service Systems with Large n.* Lecture Notes in Economics and Mathematical Systems **87**, Springer-Verlag, Berlin.

[9] Rockafellar, R.T. (1970) *Convex Analysis.* Princeton University Press, Princeton, N.J.

[10] Whitt, W. (1982) On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Prob.* **14**, 171–190.

TRAFFIC MODELING FOR HIGH-SPEED NETWORKS: THEORY VERSUS PRACTICE

WALTER WILLINGER*

Abstract. Statistical analyses of large sets of traffic measurements from working packet networks show that, from a statistical view point, traditional modeling assumptions such as Poisson packet arrivals, exponential service and Markovian structures have little to do with reality. Instead, the analyses provide convincing evidence for the presence of features in actual network traffic such as long-range dependence ("Joseph Effect"), the infinite variance syndrome ("Noah Effect") and self-similarity ("fractals"). We present some results of our traffic data analysis and discuss their implications for traffic modeling of high-speed communications systems. In particular, we point out directions in which traditional traffic models should be extended (i) to describe basic characteristics observed in measured traffic more accurately and (ii) to be more relevant for practical applications in modern telecommunications.

Key words. Infinite variance, long-range dependence, self-similarity, fractal traffic.

1. Introduction. Traffic is the driving force of communications systems, and traffic models are of crucial importance for assessing their performance. In practice, stochastic models of traffic streams are relevant to network traffic engineering and performance analysis, to the extent that they are able to predict system performance measures to a reasonable degree of accuracy. The fundamental systems, of which traffic is a major ingredient, are queueing systems. Traditional traffic models have often been devised and selected for the analytical tractability they induce in the corresponding queueing system. However, a practitioner's confidence in a given traffic model is greatly diminished if the model is only able to crudely approximate basic statistics but cannot capture visually dominant features of empirical traffic collected from a variety of working communications systems. While originally the validity and efficacy of models for modern high-speed network traffic was difficult to assess due to the unavailability of empirical data, recently very large sets of traffic measurements from working packet networks have become available (e.g., from Common Channel Signaling Networks (CCSNs) at 56 kbps, from Integrated Services Digital Networks (ISDN) at 1.5 Mbps, and from Ethernet local area network (LANs) at 10 Mbps). More importantly, statistical analyses of these enormous traffic data sets (see for example, [6], [21], [17]) have revealed features in measured network traffic that (i) have gone unnoticed by the teletraffic literature, (ii) show that from a statistical view point traditional traffic models have little in common with empirical data from modern high-speed networks, and (iii) seem to have serious implications for the design, management and control of modern telecommunications systems.

* Bellcore, 445 South Street, Room 2P-372, Morristown, NJ 07960-6438, email: walter@bellcore.com