Proceedings of the 26th Conference
on Decision and Control
Los Angeles, CA • December 1987

HOW LARGE DELAYS BUILD UP IN A GI/G/1 QUEUE

V. Anantharam

School of Electrical Engineering
Cornell University, Ithaca, NY 14853

ABSTRACT

Let W_k denote the waiting time of customer k, $k \geq 0$, in an initially empty GI/G/1 queue. Fix $a > 0$. We prove weak limit theorems describing the behavior of $\frac{W_k}{n}$, $0 \leq k \leq n$, given $W_n > na$. Let X have the distribution of the difference between the service and interarrival distributions. We consider queues for which Cramer type conditions hold for X, and queues for which X has regularly varying positive tail.

I. INTRODUCTION

In this paper we describe how large delays build up in a GI/G/1 queue. The engineering motivation for this investigation comes from the desire to design computer networks to minimize the rate of occurrence of bad events, such as buffer overflow or the occurrence of large delays. Typically, the probability of occurrence of such events needs to be determined by simulation. Knowing how such events occur can lead to fast simulation schemes. The relevant ideas are briefly discussed in Section II.

We are given an initially empty GI/G/1 queue. Customer 0 arrives at time 0 and customer k at time $A_1 + \ldots + A_k$, where A_i, $i = 1, 2, \ldots$ are i.i.d. . Let B_i, $i = 1, 2, \ldots$ be i.i.d. service times, B_{k-1} being the service time of customer k. If we let $X_k = B_k - A_k$, then X_i, $i = 1, 2, \ldots$ are i.i.d. . Let $EX_1 = -\mu$, $\mu > 0$. The waiting time W_k of customer k before receiving service is given by the equations:

$$W_0 = 0,$$

$$W_{k+1} = \max(0, W_k + X_{k-1}), \; k \geq 0.$$

Let C denote $C[0, 1]$, the space of continuous functions on the unit interval with the topology of uniform convergence, and D denote $D[0, 1]$, the space of right continuous functions with left limits on the unit interval, with the topology of uniform convergence. For each $n \geq 1$, let $w_n(\cdot) \in C$ denote the polygonal path, and $\hat{w}_n(\cdot) \in D$ denote the piecewise constant right continuous path, constructed from the points $(\frac{k}{n}, \frac{W_k}{n})$, $0 \leq k \leq n$.

Fix $a > 0$. Let P_n denote the conditional distribution $(w_n(\cdot) | W_n > na)$, and $\hat{P}_n$ the conditional distribution $(\hat{w}_n(\cdot) | W_n > na)$. P_n (resp. $\hat{P}_n$) is a probability distribution on C (resp. D).

In Theorem 1, we identify the weak limit of $\{P_n\}$ under the following moment conditions on X_1:

(C) Let $m(t) = E \exp tX_1$. There is an interval $S = (t_-, t_+)$ containing 0, such that:

$$m(t) < \infty \text{ for all } t \in S,$$

$$\frac{m'(t)}{m(t)} \to \infty \text{ as } t \to t_+,$$

$$\frac{m'(t)}{m(t)} \to -\infty \text{ as } t \to t_- .$$

This situation covers the M/M/1 queue and a large class of GI/G/1 queues with rapidly decreasing tail distributions for X_1.

In Theorem 2, we identify the weak limit of $\{\hat{P}_n\}$ under the following moment conditions on X_1:

(D) There is $q > 0$ and a slowly varying function $L(\cdot)$ such that:

$$EX_1^2 < \infty,$$

$$P(X_1 > x) \sim x^{-q} L(x) \text{ as } x \to \infty.$$

This situation covers a large class of GI/G/1 queues where X_1 has a fat positive tail.

II. MOTIVATION

The motivation for understanding how rare events occur comes from the need to speed up the estimation of the probability of occurrence of these events via simulation. This need arises in the design of computer networks to minimize the rate of occurrence of bad events such as buffer overflow or the occurrence of large delays. In a network that is performing reasonably when it is well approximated by a stationary model, these events are already relatively rare. Further, it is typically not easy to give tractable analytical formulas for the probabilities of such events.

Let (Ω, F, P) be a probability space, for example, the space of paths of a computer network, and $A \subseteq \Omega$ be a rare event, i.e. $P(A)$ is small. To estimate $P(A)$ we may observe independent samples $V_1, V_2, \ldots$, with distribution P and construct the empirical estimator

$$\overline{V}_n = \frac{1}{n} \sum_{k=1}^{n} 1(V_k \in A).$$

Clearly,

$$E_P \overline{V}_n = P(A)$$

and

$$var_P \overline{V}_n = [\frac{P(A)(1 - P(A))}{n}].$$

Suppose instead we simulate the network under a modified set of dynamics corresponding to a different probability distribution Q on (Ω, F). We observe samples $V_1^*, V_2^*, \ldots$ with distribution Q and construct the estimator

$$\overline{V}_n^* = \frac{1}{n} \sum_{k=1}^{n} \frac{dP}{dQ} \, 1(V_k^* \in A).$$

Then

$$E_Q \overline{V}_n^* = P(A)$$

and

$$var_Q \overline{V}_n^* = \frac{1}{n} (\int_A \frac{dP}{dQ} \, dP - P(A)^2).$$

If Q is chosen so that $\frac{dQ}{dP} \gg 1$ on A, the estimator $\overline{V}_n^*$ has much smaller variance than $\overline{V}_n$. In fact, if $\frac{dQ}{dP} \sim L \gg 1$ on A, we have

$$\frac{var_Q \overline{V}_n^*}{var_P \overline{V}_n} \sim \frac{[L^{-1} - P(A)]}{[1 - P(A)]}$$

$$\sim L^{-1}.$$

The above idea, called the importance sampling technique, is well known to experts in simulation, see, e.g., Bratley, et al., [4] and Siegmund, [11]. Because of the smaller variance of $\overline{V}_n^*$, simulating under Q generates acceptable confidence bounds with fewer samples. Intuitively, to maximize the speedup in simulation time, it is desirable to concentrate the measure Q around the paths in A along which the event A is most likely to occur. This explains our interest in the problem addressed in this paper.

For other work motivated by the same ideas and including striking numerical evidence of their power, see Cottrell et al., [5], and the thesis of Parekh, [10].

III. MAIN RESULTS

1. Rapidly decreasing tails

Under the moment conditions (C), $m(\cdot)$ is an analytic, strictly convex function on S, as is $\log m(\cdot)$. Note that $\frac{m'(0)}{m(0)} = -\mu$. Let $t_0 > 0$ be the unique point where $m(t_0) = 1$, and let z_0 denote $\frac{m'(t_0)}{m(t_0)}$. See Figure 1.

Theorem 1: Assume the moment conditions (C). If $a \leq z_0$, let $t(a) = 1 - \frac{a}{z_0}$ and $p_a(\cdot) \in C$ be defined by:

$$p_a(t) = 0, \qquad 0 \leq t \leq t(a),$$
$$p_a(t) = z_0(t - t(a)), \quad t(a) \leq t \leq 1.$$

If $a \geq z_0$, let $p_a(\cdot) \in C$ be defined by:

$$p_a(t) = ta, \qquad 0 \leq t \leq 1.$$

See Figure 2.

We have

$$P_n \overset{w}{\Rightarrow} \delta_{p_a},$$

CH2505-6/87/0000-1871$1.00 © 1987 IEEE

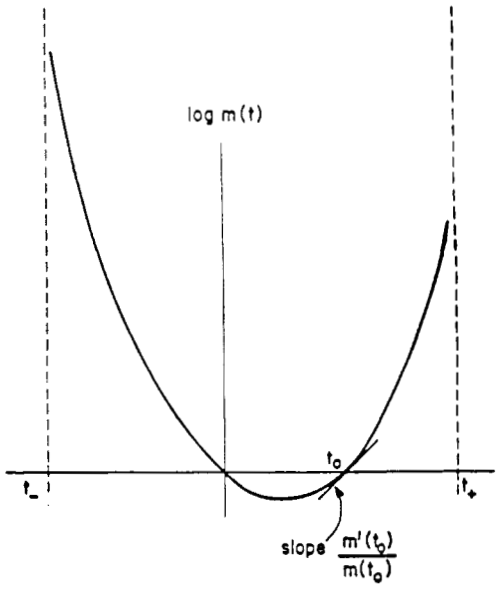

Figure 1.

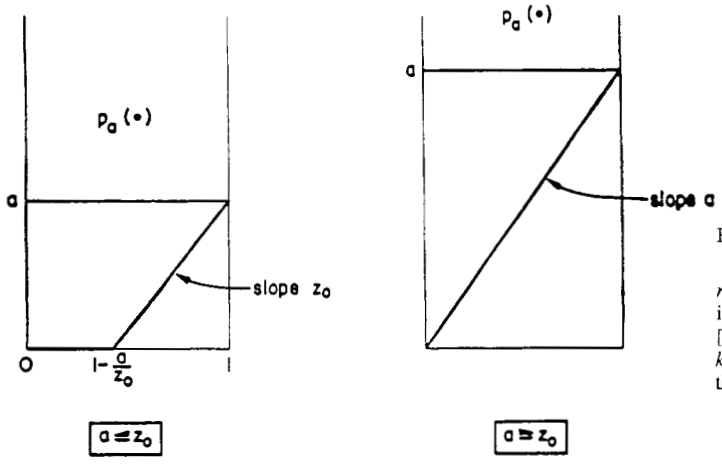

Figure 2.

where δ_{p_a} denotes the probability distribution concentrated on $p_a(\cdot)$, and $\overset{w}{\Rightarrow}$ denotes weak convergence of probabilities on C. □

For the meaning of weak convergence, see Billingsley, [2]. Heuristically, Theorem 1 states that the most likely way in which the waiting time of customer n builds up to na as follows: If $a \leq z_0$, customers k, $k \leq [nt(a)]$, encounter essentially typical behavior in the queue, after which the waiting time of customers builds up essentially linearly. If $a \geq z_0$, waiting time starts building up right away and builds up essentially lincarly.

2. Fat tails

Recall that a function $L(\cdot)$ is called slowly varying if

$$\frac{L(tx)}{L(x)} \to 1 \text{ for any } t > 0.$$

When $P(X_1 > x) \sim x^{-q}L(x)$ as $x \to \infty$, with $q > 0$ and $L(\cdot)$ slowly varying, X_1 is said to have regularly varying positive tail. For more about regularly varying distributions see Feller, Secs. VIII.8 and VIII.9, [7].

Theorem 2: Under the moment conditions (D)

$$\hat{P}_n \overset{w}{\Rightarrow} [J_T - \mu(\cdot - T)]1(T < \cdot),$$

where T is distributed on $[0, 1]$ with

$$P(T < t) = \frac{\int_t^1 (\frac{a+s\mu}{a})^{-q} ds}{\int_0^1 (\frac{a+s\mu}{a})^{-q} ds},$$

and J_T is distributed on $[a + \mu(1 - T), \infty)$ with

$$P(J_T > x) = (\frac{x}{a+\mu(1-T)})^{-q}.$$

See Figure 3.

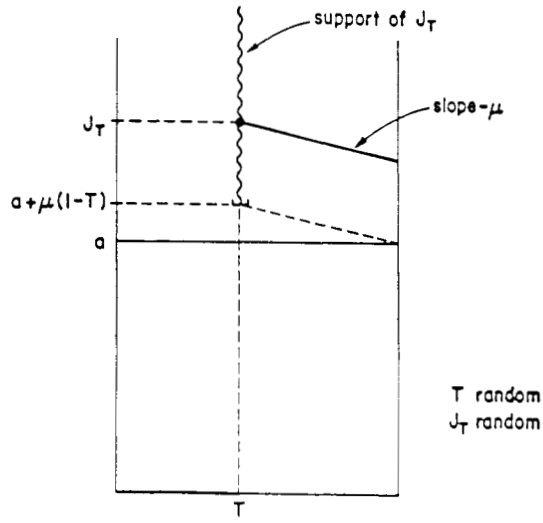

Figure 3.

Here $\overset{w}{\Rightarrow}$ denotes weak convergence on D. □

Heuristically, Theorem 2 says that if X_1 has fat tails, the reason customer n incurred a waiting time of at least na is that a single (random) customer $[kT]$ incurred an enormous waiting time. The other customers preceding customer $[kT]$ encountered typical behavior in the queue, while the customers k, $[kT] < k \leq n$, encountered typical behavior in the queue but had large waiting times because of the waiting time of customer $[kT]$.

REFERENCES

[1] H. Bergstrom, Weak Convergence of Measures, Academic Press, 1982.

[2] P. Billingsley, Convergence of Probability Measures, John Wiley and Sons, 1968.

[3] A. A. Borovkov, "Boundary value problems for random walks and large deviations in function spaces", Theory of Probability and its Applications, vol. 12, no. 4, pp. 575-595, 1967.

[4] P. Bratley, B. L. Fox and L. E. Schrage, A Guide to Simulation, Springer Verlag, 1983.

[5] M. Cottrell, J. Fort and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms", IEEE Trans. on Aut. Cont., vol. AC-28, no. 9, pp. 907-920, Sept. 1983.

[6] R. Durrett, "Conditional limit theorems for random walks with negative drift", Zeit. fur Wahr. und Ver. Geb., vol. 52, pp. 277-287, 1980.

[7] W. Feller, An Introduction to Probability Theory and its Applications, Vol. II, John Wiley and Sons, 1971.

[8] A. A. Mogulskii, "Large Deviations for Trajectories of Multidimensional Random Walks", Theory of Probability and its Applications, vol. XXI, no. 2, pp. 300-315, 1976.

[9] J. Neveu, Discrete Parameter Martingales, North Holland, 1975.

[10] S. Parekh, Doctoral Dissertation, Universitiy of California, Berkeley, 1986.

[11] D. Siegmund, "Importance sampling in the Monte Carlo study of sequential tests", Annals of Statistics, vol. 4, no. 4, pp. 673-684, 1976.