Accepted to the Journal of Applied Probability

Hitting Time in an M/G/1 Queue

Sheldon M. Ross [1]
Department of Industrial Engineering and Operations Research
University of California, Berkeley, CA 94720.

Sridhar Seshadri
Department of Statistics and Operations Research & Operations Management Area
Leonard N. Stern School of Business
New York University, NY 10012.

Abstract

We study the expected time for the work in an M/G/1 system to exceed the level x, given that it started out initially empty, and show that it can be expressed solely in terms of the Poisson arrival rate, the service time distribution and the stationary delay distribution of the M/G/1 system. We use this result to construct an efficient simulation procedure.

[1] The research of Sheldon M. Ross was supported by the National Science Foundation Grant DMI-9610046 with the University of California.

1 INTRODUCTION

For an initially empty M/G/1 system we determine the expected time until the work in the system exceeds a given value x. Let, $G(y), 0 \leq y \leq x$, denote the expected time to exceed the level x given that the work is currently $x - y$. In this notation, the expected hitting time is $G(x)$. In section 2, we derive and invert the transform of $G(y)$. We then show that $G(x)$ can be expressed solely in terms of the Poisson arrival rate, the service time and the stationary delay distributions in the M/G/1 system. Based on the formula derived in section 2, we provide an efficient simulation procedure for computing $G(x)$ in section 3. The transform of the hitting time had previously been derived by Cohen [2], section III.5.8 (but had not been inverted) using a method that is different than the one used in this paper.

2 EXPECTED HITTING TIME

The arrival process of customers to the M/G/1 queue is Poisson with rate λ, and the service times are i.i.d. rv's with finite first and second moments. The service time distribution is denoted as F and the service rate is denoted as $\mu = 1/E[S]$. We assume that the service time distribution has no probability mass at 0. The utilization of the system is expressed as $\rho = \lambda/\mu$. We assume that $\rho < 1$. Let W represent a random variable that has the same distribution as the stationary delay in the M/G/1 queue. Let S be a random variable that is independent of W and has the distribution F. The expected length of a busy period in the M/G/1 queue is given by $E[S]/(1-\rho)$, see Wolff [4]. An idle period followed by a busy period will be called a cycle. The expected length of a cycle is given by (see Ross [3] for example)

$$E[S]/(1-\rho) + 1/\lambda \quad = \quad \frac{1}{\lambda(1-\rho)}. \tag{1}$$

The M/G/1 system regenerates after the end of each cycle. Let p be the probability that the work ever exceeds x in a busy cycle. Let q be the conditional probability, given that the work is currently x, that it will exceed x at least one more time in the same busy cycle.

We shall derive expressions for p and q by setting up a differential equation. A similar equation is used later in this section to derive an expression for $G(y)$. The formulae for p and q are derived in Chapter III, section 8, page 551 of Cohen [2] using a different approach.

Theorem 2.1

$$p = \frac{\Pr\{x - S \leq W < x\}}{\Pr\{W < x\}}, \tag{2}$$

$$q = \frac{\Pr\{W < x\} - (1-\rho)}{\Pr\{W < x\}}. \tag{3}$$

Proof: Let $P(y)$ be the probability that the work does not exceed x in the busy cycle, given that the work is currently $x - y$. Thus, $P(0) = 1 - q$. For $0 \leq y < x$, we may condition on whether or not a new customer arrives within the next Δy time units to obtain

$$P(y) = (1 - \lambda\Delta y)P(y + \Delta y) + \lambda\Delta y \int_0^{y+\Delta y} P(y + \Delta y - u)dF(u) + o(\Delta y).$$

Rearranging, dividing by Δy, and taking limits as Δy goes to zero, yields the equation

$$-\frac{dP(y)}{dy} = -\lambda P(y) + \lambda \int_0^y P(y-u)dF(u), \quad 0 \leq y < x. \tag{4}$$

Let $P^*(s)$ be the Laplace Stieltjes Transform (LST) of $P(y)$. Taking transforms of both sides of equation (4) and rearranging gives

$$P^*(s) = \frac{1-q}{s - \lambda + \lambda F^*(s)}.$$

Using the known result for the LST of the stationary delay (see Cohen [2], section II.4.5, page 255) yields

$$P(y) = \frac{1-q}{1-\rho}\Pr\{0 \leq W < y\}.$$

We can directly obtain a formula for q from this equation by observing that, $\lim_{y \to x} P(y) = 1$. Therefore, (cf: equation (3))

$$1 - q = \frac{1-\rho}{\Pr\{W < x\}}. \tag{5}$$

We need another equation to determine p. Observe that $1 - p$ is the probability that the work does not exceed the level x during a busy cycle. Therefore, upon conditioning on the service time of the first customer in the busy period, we can express

$$1 - p = \int_0^x P(x-y)dF(y) \tag{6}$$

$$= \frac{1-q}{1-\rho}\int_0^x \Pr\{W < x-y\}dF(y)$$

$$= \frac{(1-q)\Pr\{W+S<x\}}{1-\rho}. \qquad (7)$$

Equations (3) and (7) together yield

$$1-p = \frac{\Pr\{W+S<x\}}{\Pr\{W<x\}}.$$

□

Our approach for obtaining $G(x)$ is a little indirect. We first derive a differential equation for $G(y)$. Second, we use this equation to determine the expected overshoot above the level x under two different conditions, namely (i) when it takes place for the first time in a cycle, and (ii) when it takes place after the first overshoot but in the same busy cycle. Finally, we finish off with obtaining the value of $G(x)$. Recall that $G(y), 0 \leq y \leq x$ is the expected time to hit or exceed the level x given that the workload is currently $x-y$. For $y < x$ we may condition on whether or not a new customer arrives within the next Δy time units to obtain

$$G(y) = \Delta y + (1-\lambda\Delta y)G(y+\Delta y) + \lambda\Delta y \int_0^{y+\Delta y} G(y+\Delta y-u)dF(u) + o(\Delta y).$$

Rearranging, dividing by Δy, and taking limits as Δy goes to zero, we get the equation

$$-\frac{dG(y)}{dy} = 1-\lambda G(y) + \lambda\int_0^y G(y-u)dF(u).$$

Denote the LST of $G(y)$ as $G^*(s)$. Taking the LST of both sides of this equation we obtain

$$-sG^*(s) + G(0) = 1/s - \lambda G^*(s) + \lambda G^*(s)F^*(s)$$

implying that

$$G^*(s) = \frac{G(0)-1/s}{s(1-\rho)}\frac{s(1-\rho)}{s-\lambda+\lambda F^*(s)}.$$

The last quantity in the right hand side of this equation is the LST of the stationary delay distribution. Let $I(.)$ stand for the identity function ($I(y)=y$), and $\circ$ for the convolution operation. Let $H(y) = \Pr\{W \leq y\}$. We can invert the transform $G^*(s)$ and write

$$G(y) = \frac{1}{1-\rho}(G(0)-I)\circ H(y). \qquad (8)$$

The value of $G(0)$ must still be determined. We proceed to do so now. Let the work in the system currently be x. Let A be the event that the work again exceeds x before the end of the present busy cycle. In our notation, $\Pr(A) = q$. Let V_x be the expected time to exceed x given that event A occurs. Let $E[\Delta]$ be the overshoot above the level x given that event A happens. Let V_0 be the expected time for the work in the system to reach the level 0 given that event A does not happen. We also know that the expected time to hit 0 unconditionally is $x/(1-\rho)$, i.e., it is equal to the expected length of a busy period which begins with an exceptional service time equal to x, see for example Wolff [4]. Then conditioning on the event A happening or not, we obtain,

$$\frac{x}{1-\rho} = qV_x + q\left(\frac{x+E[\Delta]}{1-\rho}\right) + (1-q)V_0. \tag{9}$$

Using a similar argument, $G(0)$ equals V_x if A happens, else it equals the expected time to hit 0 plus the time $G(x)$. Thus we get

$$G(0) = qV_x + (1-q)V_0 + (1-q)G(x). \tag{10}$$

Combining equations (9) and (10) we obtain the second equation for $G(0)$,

$$G(0) = \frac{(1-q)x - qE[\Delta]}{1-\rho} + (1-q)G(x). \tag{11}$$

Substituting the expression for $G(0)$ from equation (11) into equation (8) (with y set equal to x), we get

$$(1-\rho)G(x) = \left(\frac{(1-q)x - qE[\Delta]}{1-\rho} + (1-q)G(x)\right)\Pr\{W < x\} - I \circ H(x).$$

Using (3) and simplifying yields

$$E[\Delta] = \frac{(1-\rho)(x - I \circ H(x))}{\Pr\{W < x\} - (1-\rho)}. \tag{12}$$

The reader can verify that the expected overshoot is $1/\mu$ for the M/M/1 system. Because there is a probability mass at zero, the expression $I \circ H(x)$ should be interpreted as $x - \int_0^x \Pr\{W > y\}dy$.

Having obtained $E[\Delta]$ it becomes relatively simple to evaluate the expected value of the jump above x the *first* time the work hits or exceeds x during a busy cycle. Denote the expected value of the first jump as $E[\Delta_0]$. From equation (3) we can determine the average time the work is above the level x during a busy cycle, say $L(x)$, as follows. The first jump contributes an expected duration of $\frac{E[\Delta_0]}{1-\rho}$ to $L(x)$. The subsequent jumps (a geometric number of them with parameter q) contribute, $\frac{q}{1-q}\frac{E[\Delta]}{1-\rho}$. Therefore,

$$L(x) \;=\; p\left(\frac{E[\Delta_0]}{1-\rho}+\frac{q}{1-q}\frac{E[\Delta]}{1-\rho}\right). \tag{13}$$

The ratio of $L(x)$ to the length of a cycle (see (1)) should be the fraction of time the work is above the level x, and must equal by PASTA, $\Pr\{W > x\}$. Making this connection and simplifying we obtain

$$E[\Delta_0] \;=\; \frac{\Pr\{W > x\}}{\lambda p} - [x - I \circ H(x)]. \tag{14}$$

Once again the expected value of the first overshoot equals $1/\mu$ for the M/M/1 case.

Theorem 2.2

$$G(x) \;=\; \frac{1}{\lambda p(1-\rho)} - \frac{\left(\frac{\Pr\{W>x\}}{\lambda p} + I \circ H(x)\right)}{1-\rho}. \tag{15}$$

Proof: The expected number of cycles necessary to hit or exceed the level x is given by the expectation of a geometrically distributed random variable with parameter p times the average length of a cycle, i.e., $\frac{1}{\lambda p(1-\rho)}$. The value of $G(x)$ is this quantity minus the expected time to hit the level zero having hit or exceeded the level x. From equation (14) the expected time to reach level zero once having hit or exceeded the level x is $\frac{x+E[\Delta_0]}{1-\rho}$.
□

Remark: For the M/M/1 queue, the value of $G(x)$ can be explicitly given as

$$G(x) \;=\; \frac{1-\rho e^{-(\mu-\lambda)x}}{\lambda(1-\rho)^2 e^{-(\mu-\lambda)x}} - \frac{x+\frac{1}{\mu}}{1-\rho}.$$

3 SIMULATION

In this section, we present a method that employs simulation to obtain the quantities p, $\Pr\{W > x\}$, and $I \circ H(x)$ required for computing $G(x)$. As noted earlier the expression $I \circ H(x)$ should be interpreted as $x - \int_0^x \Pr\{W > y\}dy$. This can be further simplified and written as

$$I \circ H(x) = x - E[W] + E[(W-x)^+], \tag{16}$$

where $(W - x)^+$ equals $(W - x)$ if $W > x$, and equals zero otherwise. Denote the equilibrium distribution of the service time distribution as $F_e(y) = \mu \int_0^y F^c(x)dx$, where $F^c(y) = 1 - F(y)$. W can be represented as

$$W = \sum_{i=1}^{N} X_i$$

where $X_1, X_2, \ldots$ are independent random variables distributed according to F_e, and N is independent of these X_i, and is such that

$$\Pr\{N = n\} = \rho^n (1 - \rho), \quad n \geq 0.$$

When ρ is small, we can estimate $\Pr\{W > x\}$ by using

$$\Pr\{W > x\} = \rho(1 - \rho) F_e^c(x) + \rho^2 \Pr\{W > x | N \geq 2\}.$$

The term $\Pr\{W > x | N \geq 2\}$ can be estimated by simulating N conditional on it being at least 2 (which can be accomplished by generating a geometric with parameter $1 - \rho$ and then adding 1 to this generated value) and then simulating $X_1, \ldots, X_{N-1}$ according to F_e. If $\sum_{i=1}^{N-1} X_i = s$, then the estimate of $\Pr\{W > x\}$ from this run is

$$\rho(1 - \rho) F_e^c(x) + \rho^2 F_e^c(x - s).$$

The quantity $\Pr\{W + S > x\}$ can be similarly estimated. To estimate $E[(W - x)^+]$ we use a similar representation:

$$E[(W - x)^+] = \rho(1 - \rho) E[(X - x)^+] + \rho^2 E[(W - x)^+ | N \geq 2]$$

and use simulation to estimate the latter conditional expectation.

When ρ is not particularly small we recommend a different simulation procedure. With the same notation as in the preceding, let

$$M(x) = \min\{n : \sum_{i=1}^{n} X_i > x\}.$$

Since

$$\Pr\{W > x | M(x)\} = \Pr\{N \geq M(x) | M(x)\} = \rho^{M(x)}$$

it follows that $\rho^{M(x)}$ is an unbiased estimator of $\Pr\{W > x\}$. By then generating a service time random variable S, we can use the same data to estimate $\Pr\{W + S > x\}$ by $\rho^{M(x-S)}$, where $M(y)$ should be taken to equal 0 when y is negative.

To estimate $E[(W-x)^+]$, let

$$O = \sum_{i=1}^{M(x)} X_i - x.$$

Using the lack of memory property of N gives

$$\begin{aligned} E[(W-x)^+] &= E[W-x|N \geq M(x)]\Pr\{W > x\} \\ &= E[O + E[W]|N \geq M(x)]\Pr\{W > x\} \\ &= \Pr\{W > x\}(E[W] + E[O|N \geq M(x)]). \end{aligned}$$

To estimate $E[O|N \geq M(x)]$ perform n simulation runs, where a simulation run generates the value of $M(x)$. Imagine that we had also generated the random variable N in each run, and let I_i be the indicator for the event that $N \geq M(x)$ in the i^{th} run. Then, with O_i being the value of O in the i^{th} run, it follows from the strong law of large numbers that

$$\frac{O_1 I_1 + \ldots + O_n I_n}{I_1 + \ldots + I_n}$$

is a consistent estimator of $E[O|N \geq M(x)]$. However, since N is independent of $M(x)$, it is intuitively clear that replacing I_i in the preceding by its conditional mean given $M_i(x)$, the value of $M(x)$ obtained in the i^{th} run, should result in an improved estimator. As a result, we suggest the estimator

$$\frac{\sum_{i=1}^n O_i \rho^{M_i(x)}}{\sum_{i=1}^n \rho^{M_i(x)}}.$$

We now argue that this too is a consistent estimator of $E[O|N \geq M(x)]$. To see this, note that dividing the numerator and denominator by n shows that, with probability 1,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n O_i \rho^{M_i(x)}}{\sum_{i=1}^n \rho^{M_i(x)}} = \frac{E[O\rho^{M(x)}]}{E[\rho^{M(x)}]}$$

whereas

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n O_i I_i}{\sum_{i=1}^n I_i} = \frac{E[OI]}{E[I]}.$$

Since $E[\rho^{M(x)}] = E[I]$, consistency will follows if we can show that

$$E[OI] = E[O\rho^{M(x)}].$$

This is shown as follows:

$$\begin{aligned} E[OI] &= E[E[OI|M(x)]] \\ &= E[E[O|M(x)]E[I|M(x)]] \quad \text{since given } M(x),\ I \text{ and } O \text{ are independent} \\ &= E[\rho^{M(x)}E[O|M(x)]] \\ &= E[E[O\rho^{M(x)}|M(x)]] \\ &= E[O\rho^{M(x)}] \end{aligned}$$

thus showing that our suggested estimator is consistent.

It is intuitive that the quality of our suggested estimator should be roughly the same for all x; what does vary with x is the time to generate a simulation run - that is, the time to generate $M(x)$. However, since $M(x)$ is equal to 1 plus the number of renewals by time x of the renewal process having interarrival distribution F_e, it follows from the elementary renewal theorem that

$$E[M(x)] \approx \frac{x}{E[X]} = \frac{2E[S]x}{E[S^2]}.$$

Therefore, the computational effort grows linearly in x. Please see Asmussen and Rubinstein [1] for a discussion on the computational complexity of simulating rare events. Simulation results obtained using our suggested procedure are shown in Table 1, along with those computed using the direct approach of simulating the M/G/1 queue. The simulations were coded in $f77$, and executed on a SUN SPARC station. As anticipated, for the same number of runs, the cpu time grows exponentially with x when simulating the M/G/1 queue, whereas, only linearly when using the suggested procedure.

Acknowledgements

We gratefully acknowledge the comments of an anonymous referee, whose careful scrutiny helped improve the presentation of this paper considerably. We thank the referee for spotting a discrepancy in the hitting time formula when x goes to zero. This resulted in the modification of equations (14) and (15).

Service Time Distribution	Server Utilization	x	n	Direct Simulation of M/G/1 Queue			Suggested Simul. Procedure	
				$G(x)$	cpu (sec)	Std. Err.	$G(x)$	cpu (sec)
Exponential with rate 1.1	0.909	0.5	50,000	1.697	0.35	0.0077	1.686	0.46
		2.0	50,000	5.771	1.77	0.0252	5.677	0.77
		16.0	50,000	297.314	32.19	1.2202	298.969	3.58
Sum of 5 Exp. with rates 3.66, 10.11, 4.7, 6.7 and 6.0.	0.900	0.5	500,000	1.183	15.87	0.0017	1.180	21.8
		2.0	500,000	6.665	80.37	0.0088	6.60	58.11
		16.0	500,000	924.611	11100.10	1.2477	923.145	219.20

Table 1: Comparison of Simulation Procedure with Direct Simulation of Queue

References

[1] S. Asmussen and R. Y. Rubinstein. Steady State Rare Events Simulation in Queueing Models and its Complexity Properties. (ed.) J. H. Dshalalow. *Advances in Queueing: Theory, Methods, and Open Problems*, CRC Press, Boca Raton, FL, (1995)

[2] Cohen, J. W. *The Single Server Queue.* Revised edition, North-Holland Pub. Co., Amsterdam, 1982.

[3] Ross, S. M. *Stochastic Processes.* John Wiley & Sons, New York, NY, 1983.

[4] Wolff, R. W. *Stochastic Modeling and the Theory of Queues.* Prentice Hall, Englewood Cliffs, NJ, 1989.