Queueing Systems 8 (1991) 105–110 

*SHORT COMMUNICATION*

# SETTLING TIME BOUNDS FOR $M/G/1$ QUEUES

Arif MERCHANT *

*Department of Computer Science, Stanford University, Stanford, CA 94305, USA*

Received 30 October 1989; revised 3 August 1990

This paper addresses the question of how long it takes for an $M/G/1$ queue, starting empty, to approach steady state. A coupling technique is used to derive bounds on the variation distance between the distribution of number in the system at time $t$ and its stationary disribution. The bounds are valid for all $t$.

**Keywords:** Settling times, $M/G/1$ queues, coupling, relaxation times.

## 1. Introduction and a basic result

The $z$-transform solution for the number in the system of an $M/G/1$ queue has long been known [4,7]. We are interested in the question of settling times: how long does it take, for a queue starting empty, to approach steady state? One common approach to this problem is to consider the asymptotic behavior of the number in the system as the time $t \to \infty$; this approach is frequently called the relaxation time approximation (see Cohen [2]). We use instead the *coupling* technique to relate the settling time to the time for a stationary queue to empty. This allows us to bound the variation distance between the distribution of number in the system at time $t$ and its stationary distribution in terms of the arrival rate and the moments of the service time. The bounds obtained are valid for all $t$.

Let $Z$ be a Markov process on the state space $S$ with the transition function $P$. Assume that $Z$ has a unique stationary distribution $\pi$ and let $\pi_t$ denote the distribution of $Z$ at time $t$. We define a coupling for $Z$ as a process $(X, Y)$ on $S \times S$ with a random stopping time $T$, called the *coupling time* such that:

(1) $X$ is the Markov process with transition function $P$ and initial distribution $\pi_0$.

* This research was supported in part by a grant from the AT&T Foundation and NSF grant DCR-8351757.

(2) $Y$ is the Markov process with transition function $P$ and initial distribution $\pi$.
(3) $X_t = Y_t$ for $t \geqslant T$.
Note that $X$ and $Y$ need not be independent.

The following theorem bounds the *variation distance* $\| \pi_t - \pi \|$ between $\pi_t$, the distribution of $Z$ at time $t$ and the stationary distribution $\pi$ in terms of the tail of the distribution of the coupling time $T$.

THEOREM 1
(*Coupling inequality*)

$$\begin{aligned} \| \pi_t - \pi \| &= \sup_{A \subseteq S} | \pi_t(A) - \pi(A) | \\ &\leqslant Pr[T > t]. \end{aligned}$$

*Proof*

For any $t \geqslant 0$, $Pr[X_t \in A,\ T \leqslant t] = Pr[Y_t \in A,\ T \leqslant t]$, therefore,

$$\begin{aligned} | Pr[X_t \in A] - Pr[Y_t \in A] | &= | Pr[X_t \in A,\ T > t] - Pr[Y_t \in A,\ T > t] | \\ &\leqslant \max( Pr[X_t \in A,\ T > t],\ Pr[Y_t \in A,\ T > t]) \\ &\leqslant Pr[T > t]. \quad \square \end{aligned}$$

The above proof is a slightly modified version of that given in Thorisson [8] for discrete time stochastic processes.

Let $Z_t = (Q_t, L_t)$ be the state vector for an $M/G/1$ queue, where $Q_t$ is the number in the system at time $t$ and $L_t$ is the service time already received by the customer in service at $t$. Let $Z_0 = (0, 0)$. The process $Z = \{ Z_t\colon t \geqslant 0 \}$ is a Markov process. Let $\lambda$ be the arrival rate for customers and $\mu$ be the rate of service; for stability we assume $\lambda < \mu$. Let the distribution of $Z_t$ be denoted $\pi_t$ and the stationary distribution be $\pi$.

Define a coupling for $Z_t$ as follows:
(1) $X_t = Z_t$ for $t \geqslant 0$.
(2) $Y_t = (Q_t^*, L_t^*)$ is the state vector for the queue starting in the stationary distribution for $X$ and with the same sequence of arrivals and service times as $X$. Clearly, $Y_t$ is a Markov process with the same transition probability function as $X$.
(3) Let $T = \inf\{ t\colon Y_t = (0, 0) \}$. Clearly $Pr[T < \infty] = 1$.

Since $Q_0^* \geqslant Q_0 = 0$ and $Y$ has the same sequence of arrivals and service times as $X$, $Q_t^* \geqslant Q_t$ for $t \geqslant 0$, which implies that $X_t = Y_t$ for $t \geqslant T$.

We may therefore apply the coupling inequality:

$$\| \pi_t - \pi \| \leqslant Pr[T > t]. \tag{1}$$

Now let $V_t = (Q_t', L_t')$ be the state vector for the queue with $V_0 = (Q_0^*, 0)$ and the same arrivals and service times as $Y$ including the service times for the customers in the system at $t=0$. Let $T' = \inf\{t\colon V_t = (0, 0)\}$. A little reflection convinces us that $T' \geqslant T$, so that

$$\| \pi_t - \pi \| \leqslant Pr[T' > t]. \tag{2}$$

In the next section, we shall derive bounds on the tail of the distribution of $T'$.

## 2. The distribution of $T'$

In this section, we examine $V_t$ at the times of customer departures; since $L'_t = 0$ for these points, we shall simply denote the state by $Q'_t$.

Let

$\tau_i$ = time for $V$ to "move" from $i$ to $i-1$ for $i > 0$.

Therefore,

$$T' = \tau_1 + \tau_2 + \cdots + \tau_{Q_0^*}. \tag{3}$$

Since $\tau_1, \tau_2, \ldots$ are clearly i.i.d., we may apply standard results for random sums. We already know the distribution of $Q_0^*$, since this is the stationary distribution of number in the system, so it only remains to find the distribution of $\tau_1$.

For $\tau_1$ we have

$$\tau_1 = S + \tau'_1 + \tau'_2 + \cdots + \tau'_{N(S)}, \tag{4}$$

where $S$ is the service time for a customer, $N(S)$ is the number of arrivals in this service time, and $\tau'_1, \tau'_2, \ldots$ are i.i.d. and distributed as $\tau_1$.

Let $\Phi(w) = E[e^{-w\tau_1}]$ be the Laplace transform of the distribution of $\tau_1$. Clearly, $\Phi(w)$ exists for $\mathrm{Re}(w) \geqslant 0$. Then from eq. (4), we have

$$\begin{aligned}\Phi(w) &= E\left[\mathrm{e}^{-wS}\Phi(w)^{N(S)}\right]\\ &= E\left[E\left[\mathrm{e}^{-wS}\Phi(w)^{N(S)} \mid S\right]\right]\\ &= E\left[\mathrm{e}^{-wS}\,\mathrm{e}^{\lambda S(\Phi(w)-1)}\right] \quad \text{since } N(S) \text{ is Poisson with parameter } \lambda S\\ &= E\left[\mathrm{e}^{-S(w-\lambda\Phi(w)+\lambda)}\right]\\ &= G^*(w-\lambda\Phi(w)+\lambda),\end{aligned} \tag{5}$$

where $G^*(w) = E[\mathrm{e}^{-wS}]$ is the Laplace transform of the service time distribution.

While it is not clear how to find $\Phi(w)$ from this equation, we may readily find the moments of the distribution by differentiating:

$$\begin{aligned}E[\tau_1] &= -\Phi'(0) = (1-\lambda\Phi'(0))E[S]\\ &= ES/(1-\lambda ES),\end{aligned} \tag{6}$$

$$\begin{aligned}E[\tau_1^2] &= \Phi''(0) = (1-\lambda\Phi'(0))^2E[S^2] + \lambda\Phi''(0)E[S]\\ &= E[S^2]/(1-\lambda E[S])^3.\end{aligned} \tag{7}$$

To make further calculation compact, we rewrite these in terms of

$$\rho_i = \lambda^i E[S^i],$$

so that

$$E[\tau_1] = \frac{1}{\lambda} \cdot \frac{\rho_1}{1-\rho_1} \tag{8}$$

$$\sigma_{\tau_1}^2 = E[\tau_1^2] - E[\tau_1]^2 = \frac{1}{\lambda^2} \cdot \frac{\rho_2 - \rho_1^2 + \rho_1^3}{(1-\rho_1)^3}. \tag{9}$$

For the distribution of $Q_0^*$ we have the Pollaczek–Khinchin $z$-transform equation (see, for instance, Kleinrock [6]):

$$Q(z) = \frac{G^*(\lambda - \lambda z)(1-\rho_1)(1-z)}{G^*(\lambda - \lambda z) - z}. \tag{10}$$

The moments may be found, again, by differentiating:

$$E[Q_0^*] = Q'(1) = \rho_1 + \frac{\rho_2}{2(1-\rho_1)}, \tag{11}$$

$$\begin{aligned}\sigma_{Q_0^*}^2 &= Q''(1) + Q'(1) - Q'(1)^2 \\ &= \frac{\rho_3}{3(1-\rho_1)} + \frac{\rho_2^2}{4(1-\rho_1)^2} - \frac{(2\rho_1 - 3)\rho_2}{2(1-\rho_1)} + \rho_1(1-\rho_1).\end{aligned} \tag{12}$$

From eqs. (3) and (8)–(12) and standard results for the mean and variance of random sums (see Karlin and Taylor [3]) we have,

$$\begin{aligned}E[T'] &= E[\tau_1] \cdot E[Q_0^*] \\ &= \frac{1}{\lambda} \cdot \left( \frac{\rho_1^2}{1-\rho_1} + \frac{\rho_1 \rho_2}{2(1-\rho_1)^2} \right),\end{aligned} \tag{13}$$

$$\begin{aligned}\sigma_{T'}^2 &= E[\tau_1]^2 \sigma_{Q_0^*}^2 + E[Q_0^*] \sigma_{\tau_1}^2 \\ &= \frac{1}{\lambda^2} \cdot \{4\rho_1^2(1-\rho_1)\rho_3 + 3(2+\rho_1^2)\rho_2^2 \\ &\quad + 12\rho_1(1-\rho_1)(1+\rho_1-\rho_1^2)\rho_2 - 12\rho_1^4(1-\rho_1)^2\}\{12(1-\rho_1)^4\}^{-1}.\end{aligned} \tag{14}$$

Armed with these moments, we may then apply Chebychev's inequality to bound the tail of the distribution of $T'$:

$$Pr\left[(T' - E[T'])^2 > k^2 \sigma_{T'}^2\right] \leqslant 1/k^2$$

and for $t > E[T']$,

$$\| \pi_t - \pi \| \leqslant Pr[T' > t] \leqslant \sigma_{T'}^2 / (t - E[T'])^2. \tag{15}$$

*Example: M/M/1 queue*

For an exponential service time distribution, the Laplace transform is

$$G^*(w) = \frac{\mu}{w+\mu} \quad \text{for } \mathrm{Re}(w) > -\mu. \tag{16}$$

This gives us

$$\rho_i = \lambda^i E[S^i] = i!(\lambda/\mu)^i = i!\rho^i$$

in terms of the traffic intensity $\rho = \lambda/\mu$. Substituting in eqs. (13)–(14),

$$E[T'] = \frac{1}{\lambda} \cdot \left(\frac{\rho}{1-\rho}\right)^2$$

and

$$\sigma_{T'}^2 = \frac{1}{\lambda^2} \cdot \frac{\rho^3(2+\rho)}{(1-\rho)^4},$$

which we may use in the Chebychev inequality. However, the $M/M/1$ case can be made to yield better bounds, as we show in the next section.

## 3. Exponential bounds

When eq. (5) can be solved explicitly for $\Phi(w)$, it is possible to find stronger bounds for the variation distance, as we shall see for the $M/M/1$ case.

Substituting the Laplace transform for the service time distribution in eq. (5) we have

$$\Phi(w) = \frac{\mu}{w - \lambda\Phi(w) + \lambda + \mu} \quad \mathrm{Re}(w) \geqslant 0,\ \mathrm{Re}(w - \lambda\Phi(w) + \lambda + \mu) > 0.$$

Solving the quadratic equation for $\Phi(w)$, and selecting the negative sign to make $\Phi(0) = 1$,

$$\Phi(w) = \frac{\lambda + \mu + w - \sqrt{(\lambda + \mu + w)^2 - 4\lambda\mu}}{2\lambda} \quad \mathrm{Re}(w) \geqslant 0. \tag{17}$$

$\Phi(w)$ is analytic everywhere in the complex plane cut along the segment $[w_1, w_2]$, where $w_1$ and $w_2$ are the two (negative) roots of $(\lambda + \mu + w)^2 = 4\lambda\mu$.

For the stationary queue length distribution, we have from eq. (10)

$$Q(z) = \frac{\mu - \lambda}{\mu - \lambda z}.$$

Let $T^*(w) = E[e^{-wT'}]$ be the Laplace transform of the distribution of $T'$. Using eq. (3),

$$\begin{aligned} T^*(w) &= Q(\Phi(w)) \\ &= \frac{2(\mu - \lambda)}{\mu - \lambda - w + \sqrt{(\lambda + \mu + w)^2 - 4\lambda\mu}}. \end{aligned} \tag{18}$$

To bound the tail of the distribution, we show a Chernoff-type inequality [1]:

$$Pr[T' > t] = \int_{y=t}^{\infty} \mathrm{d}F(y) \quad F \text{ is the c.d.f. of } T'$$

$$\leqslant \int_{y=t}^{\infty} \mathrm{e}^{-(y-t)w_0} \,\mathrm{d}F(y) \quad \text{for } w_0 < 0$$

$$\leqslant \mathrm{e}^{w_0 t} T^*(w_0) \quad \text{for } w_0 < 0.$$

Since $T^*(w)$ is analytic everywhere in the complex plane cut along the segment $[w_1, w_2]$, we may allow $w_0 \downarrow w_2 = -(\sqrt{\mu} - \sqrt{\lambda})^2$, yielding

$$\| \pi_t - \pi \| \leqslant Pr[T' > t] \leqslant \left(\mathrm{e}^{-(\sqrt{\mu}-\sqrt{\lambda})^2 t}\right)\left(1 + \sqrt{\lambda/\mu}\right), \tag{19}$$

which gives us an exponentially decreasing bound for $\| \pi_t - \pi \|$.

In the case of $M/G/1$ queues, if $G^*(w)$ exists for some $w < 0$, then it can easily be shown that $\Phi(w_0)$, and therefore $T^*(w_0)$ exists for some $w_0 < 0$ (see Kingman [5], lemma 3). Thus a Chernoff-type inequality holds in this case.

For the $G/G/1$ system, Cohen [2] shows that asymptotically, as $t \to \infty$, various parameters such as the expected workload approach their limiting values exponentially fast, assuming that the Laplace transforms of the inter-arrival time distributions and the service time distribution are analytic in a complex half-plane which includes the axis $\mathrm{Re}(w) = 0$ in its interior. This indicates that it might be possible to derive a Chernoff-type bound in this case too. However, while it is quite simple to set up a coupling for the $G/G/1$ system similar to that in section 1, it is not obvious how the distribution of the coupling time may be found.

## Acknowledgement

We thank Anil Gangolli for introducing us to the subject of couplings, and for other help with this work.

## References

[1] H. Chernoff, A measure of asymptotic efficiency for tests based on the sum of observations, Ann. Math. Stat. 23 (1952) 493–509.
[2] J.W. Cohen, *The Single Server Queue*, 2nd ed. (North-Holland, 1982).
[3] S. Karlin and H.M. Taylor, *A First Course in Stochastic Processes* (Academic Press, 1975).
[4] A.Y. Khinchin, Mathematical theory of stationary queues, Mat. Sbornik 39 (4) (1932) 73–84.
[5] J.F.C. Kingman, On queues in which customers are served in random order, Proc. Cambridge Phil. Soc. 58 (1962) 79–91.
[6] L. Kleinrock, *Queueing Systems*, vol. 1 (Wiley, 1975).
[7] L. Takács, *Introduction to the Theory of Queues* (Oxford University Press, 1962).
[8] H. Thorisson, On maximal and distributional coupling, Ann. Prob. 14 (1986) 873–876.