Probability in the Engineering and Informational Sciences, **18**, 2004, 103–110. Printed in the U.S.A.

STABILITY CONDITIONS FOR GATED $M/G/\infty$ QUEUES

DIMITRA PINOTSI AND MICHAEL A. ZAZANIS

Department of Statistics
Athens University of Economics and Business
Athens 104 34, Greece
E-mail: zazanis@aueb.gr

The question of stability for the $M/G/\infty$ queue with gated service is investigated using a Foster–Lyapunov drift criterion. The necessary and sufficient condition for positive recurrence is shown to be the finiteness of the first moment of the service time distribution, thus weakening the stability condition given in Browne et al. [3].

1. THE GATED, EXHAUSTIVE, PARALLEL SERVICE SYSTEM

We consider the model described in Browne, Coffman, Gilbert, and Wright [3]. Customers arrive at an infinite server system in a Poisson stream at rate λ per unit time. Their service requirements are assumed to be independent and identically distributed (i.i.d.) random variables $\{\sigma_i\}_{i=1,2,\ldots}$, with common distribution $G(x) = P(\sigma_1 \leq x)$. The servers, working in parallel, provide gated and exhaustive service to the queue. Service is *gated* in that customers are served in stages as follows. At the beginning of a service period, which we call the *stage*, the gate opens to admit waiting customers that are transferred from the queue to the servers, and then the gate closes. The servers then begin serving all of these customers in parallel (we assume that there is an infinite number of servers available). The stage ends when the service of all customers admitted is complete. A new stage begins immediately by admitting to service all customers that have accumulated during the first stage in front of the gate. In the case where there where no arrivals during the first stage, we assume that the system waits until the first customer arrives. At this point, the gate opens and the customer is admitted immediately to service, thus initiating a service stage with a single customer. In a

© 2004 Cambridge University Press 0269-9648/04 $16.00

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press

possible variation to the above model, if at the end of a service stage there are no customers waiting in front of the gate, then the servers take a vacation (of random length), after which they return and a new service stage starts by admitting the customers that have meanwhile accumulated in front of the gate. (Service is *exhaustive* in that no vacation period ever begins as long as there are customers waiting in line.)

Denote the number of customers served in the nth stage by K_n and the duration of the nth stage by Y_n. Then it is not hard to see [3] that $(K_n)_{n\in\mathbb{N}}$ is a Markov chain with discrete state space (the nonnegative integers), whereas $(Y_n)_{n\in\mathbb{N}}$ also constitutes a Markov chain with continuous state space (the nonnegative reals). An analysis of these Markov chains is given in [3,4], together with approximations to the stationary distribution of these chains in a number of cases. A heavy-traffic analysis of such a system has been carried out in Tan and Knessl [8]. In the above-cited articles, however, the stability issue for this system has not been settled completely. Other queuing systems with gated mechanisms have been considered in [2,5,7].

Because the model in question has an infinite number of servers, one expects that there would be no need for a stability condition analogous to that of the single-server queue. In fact, if we assume that the service time distribution has bounded support and set $T := \inf\{x : G(x) = 1\} < \infty$, then the duration of all service stages is bounded above by T and this guarantees stability. However, if the support of the service time distribution is not bounded, then arbitrarily large service times are possible and the stability of the system is no longer obvious. If this were the ordinary $M/G/\infty$ queue, then the necessary and sufficient condition for the existence of steady state would be the finiteness of the mean service time: $E\sigma < \infty$. In fact, since the performance of this system in terms of delays is worse than that of the corresponding $M/G/\infty$ queue because of the existence of the gating mechanism, one expects that the finiteness of the expected service time is a necessary condition for stability. In this article, we will show, using a Foster–Lyapunov drift criterion, that it is both necessary and sufficient.

To see that this is not a trivial question, let us envision the following scenario described in different terms in Browne et al. [3]. Set $M_k := \max(\sigma_1, \sigma_2, \ldots, \sigma_k)$ and let G^k denote the distribution of M_k, where we have, of course, that $G^k(x) = (G(x))^k$ for all $x \geq 0$. The duration of a service stage is the maximum of all service times of customers served during this stage, and, thus, if we assume that the first stage serves k_1 customers, its duration, Y_1, has distribution G^{k_1}. If during Y_1 an unusually large number of customers, $k_2 \gg k_1$ arrives as a result of statistical fluctuations, the second stage, Y_2, will be distributed according to G^{k_2} and, thus, be very likely much larger than Y_1. One can imagine such a process escalating, particularly if G is heavy tailed. Browne et al. [3] have shown that *this escalation is not possible, provided that the service time distribution G has finite second moment*, thus establishing stability under this condition. We show that the finiteness of the first moment of G is, in fact, not only necessary but also sufficient in order to establish the stability of this system.

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press

2. MARKOV CHAIN OF SERVICE STAGES FOR LATTICE SERVICE TIME DISTRIBUTIONS

In order to establish the stability of the system, we will examine the Markov chain of service stage lengths, $(Y_n)_{n=1,2,\ldots}$. In fact, in this section we will simplify our task by assuming that the service time distribution G is lattice with lattice size $\delta > 0$ (i.e., that all service times are integer multiples of δ). In particular, we will set $g_j := P(\sigma = j\delta)$, $j = 1,2,\ldots$. An immediate consequence of this assumption is, of course, that the duration of the nth stage, Y_n, is also an integer multiple of δ. Thus, (Y_n) also becomes a discrete state space Markov chain. For simplicity, we will examine instead (Φ_n), where $\Phi_n := Y_n/\delta$, a Markov chain with state space the support of σ/δ (i.e., the subset of $\mathbb{N}$ defined as $\mathbb{S} = \{j \in \mathbb{N} : g_j > 0\}$).

It is easy to see that (Φ_n) is an irreducible chain on $\mathbb{S}$. Thus, if $\mathbb{S}$ is finite, it must be positive recurrent and, hence, the system must be stable since both the stage lengths and the number of customers waiting in the queue are tight. Thus, the only case of interest is when $\mathbb{S}$ is not finite.

The transition matrix for (Φ_n) can be evaluated by first computing the conditional probability that the $(n+1)$st stage serves $K_{n+1} = m$ customers, given the length of the nth stage. This amounts to computing the probability of m Poisson arrivals, taking also into consideration that in the case where we have zero arrivals during the service stage, we wait until the first arrival. Thus,

$$P(K_{n+1} = m | Y_n = i\delta) = \begin{cases} \dfrac{1}{m!}(\lambda i\delta)^m e^{-\lambda i\delta} & \text{for } m = 2,3,\ldots, \\ (\lambda i\delta + 1)e^{-\lambda i\delta} & \text{for } m = 1. \end{cases}$$

Also, recalling that $g_j := P(\sigma = j\delta)$ and $G_j := P(\sigma \le j\delta)$, we have

$$\begin{aligned} P(Y_{n+1} = j\delta | K_{n+1} = m) &= m g_j G_{j-1}^{m-1} + \frac{(m-1)m}{2} g_j^2 G_{j-1}^{m-2} + \cdots + \frac{m!}{m!} g_j^m \\ &= \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l}. \end{aligned}$$

Thus, if we denote the transition probability matrix for the Markov chain (Φ_n) by $P_{ij} := P(\Phi_{n+1} = j | \Phi_n = i)$, combining the above we obtain

$$P_{ij} = (\lambda i\delta + 1)e^{-\lambda i\delta} g_j + \sum_{m=2}^{\infty} \frac{(\lambda i\delta)^m}{m!} e^{-\lambda i\delta} \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l}.$$

Setting $\beta := \lambda\delta$, the above expression can also be written as

$$\begin{aligned} P_{ij} &= (i\beta + 1)e^{-i\beta} g_j + \sum_{m=2}^{\infty} \frac{(i\beta)^m}{m!} e^{-i\beta} \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l} \\ &= e^{-i\beta} g_j + \sum_{m=1}^{\infty} \frac{(i\beta)^m}{m!} e^{-i\beta} \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l} \\ &= e^{-i\beta} g_j + \sum_{m=0}^{\infty} \frac{(i\beta)^m}{m!} e^{-i\beta} \sum_{l=0}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l} - \sum_{m=0}^{\infty} \frac{(i\beta)^m}{m!} G_{j-1}^m e^{-i\beta}. \end{aligned} \tag{1}$$

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press

The above double sum can be simplified considerably as follows:

$$\sum_{m=0}^{\infty} \frac{(i\beta)^m}{m!} e^{-i\beta} \sum_{l=0}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l} = \sum_{l=0}^{m} \left\{ \sum_{m=l}^{\infty} G_{j-1}^{m-l} (i\beta)^m \frac{m!}{l!(m-l)!} \frac{1}{m!} \right\} g_j^l e^{-i\beta}$$

$$= \sum_{l=0}^{m} \frac{(i\beta)^l}{l!} g_j^l e^{-i\beta} e^{i\beta G_{j-1}}$$

$$= e^{i\beta g_j - i\beta + i\beta G_{j-1}}$$

$$= e^{i\beta G_j - i\beta}, \tag{2}$$

where, in the last equation, we have used the fact that $G_j = (G_{j-1} + g_j)$. Thus, from (1) and (2), using the notation $\bar{G}_j := 1 - G_j = \sum_{k=j+1}^{\infty} g_k$, the transition probability matrix can be written as

$$P_{ij} = e^{-i\beta\bar{G}_j} - e^{-i\beta\bar{G}_{j-1}} + e^{-i\beta} g_j, \qquad i, j = 1,2,3,\ldots \tag{3}$$

We also point out, for future reference, the following inequality satisfied by the transition probabilities:

$$P_{ij} = e^{-i\beta\bar{G}_j} - e^{-i\beta\bar{G}_{j-1}} + e^{-i\beta} g_j \le i\beta e^{-i\beta\bar{G}_j} g_j + e^{-i\beta} g_j. \tag{4}$$

The above is easy to see, if we note that from the mean value theorem we have

$$e^{-i\beta\bar{G}_j} - e^{-i\beta\bar{G}_{j-1}} = \int_{i\beta\bar{G}_j}^{i\beta\bar{G}_{j-1}} e^{-x}\, dx \le i\beta(\bar{G}_{j-1} - \bar{G}_j) e^{-i\beta\bar{G}_j}$$

and remember that $\bar{G}_{j-1} - \bar{G}_j = g_j$.

3. POSITIVE RECURRENCE, FINITE MEAN STAGE LENGTH, AND STABILITY

In order to establish the stability of the system, we first have to show the positive recurrence of the Markov chain (Φ_n). This problem is not easy to tackle directly because the stationary equations for (Φ_n),

$$\pi_j = \sum_{i=1}^{\infty} \pi_i P_{ij}, \qquad j = 1,2,\ldots, \tag{5}$$

turn out to be intractable. However, it is possible to address this problem using the classical Foster criterion for positive recurrence (see Asmussen [1]). Nonetheless, this is not sufficient for our purposes, as the positive recurrence of (Φ_n) (and hence of (Y_n)) is a necessary but not a sufficient condition for the stability of our system.

To clarify this point, consider an analogous situation in a simpler setting. Suppose that (X_n) is a zero drift simple random walk (with steps ± 1) on $\mathbb{Z}$. Suppose that

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press

$X_0 = 0$ with probability (w.p.) 1 and define $T_0 = 0$ and $T_n = \inf\{k > T_{n-1} : X_k = 0\}$. Thus, (T_n) are the corresponding epochs of return to zero. If we denote by $Y_n := T_n - T_{n-1}$ the corresponding durations of the excursions from state 0, then, clearly, (Y_n) is an i.i.d. sequence of random variables (with infinite mean). Viewed as a Markov chain, (Y_n) *is positive recurrent* even though (X_n) is not.

Returning to our system, let us denote by $X_t = (U_t, V_t)$ the process whose first component is the number of customers waiting behind the gate and the second component is the number of customers in service at time t. Also, let $\{T_n\}$ denote the point process of the epochs when the gate opens and K_n denote the number of customers admitted into the system at epoch T_n. Then, it should be clear that (X_t) is a semi-regenerative process (e.g., see [1]) with respect to the Markov renewal process (T_n, K_n). Thus, it is easy to see that the system will be stable (in the sense that it will possess a stationary distribution) if and only if in the *synchronous version* of the process (where time $t = 0$ coincides with a typical point T_0 of the Markov renewal process in the sense that K_0 is distributed according to the stationary distribution of the embedded Markov chain), $ET_1 - T_0 < \infty$. This condition, however, translates to the requirement of positive recurrence for the Markov chain (Φ_n) and the finiteness of the mean of its stationary distribution (5); that is,

$$\sum_{i=1}^{\infty} i\pi_i < \infty. \tag{6}$$

4. STABILITY CONDITION

The basic tool for establishing not only the positive recurrence of (Φ_n) but also the finiteness of the first moment of the stationary distribution (6) is the following generalization of the classic criterion of Foster. (This is Proposition 2.9 of Tweedie [10]. See also Meyn and Tweedie [6] for a comprehensive account of Foster–Lyapunov drift criteria in general state space Markov chains.)

THEOREM 1 (FOSTER–LYAPUNOV CRITERION): *Suppose that a Markov chain with countable state space, say $\mathbb{Z}_+$, and transition probability matrix P_{ij} is irreducible and let N be a given natural number. If $V: \mathbb{Z}_+ \to \mathbb{R}_+$ is a nonnegative function and $f: \mathbb{Z}_+ \to \mathbb{R}_+$ a strictly positive function on the state space such that*

$$\sum_{j \in E} P_{ij} V(j) \le b, \qquad i \le N, \tag{7}$$

$$\sum_{j \in E} P_{ij} V(j) \le V(i) - f(i), \qquad i > N, \tag{8}$$

where $b > 0$, then the Markov chain is positive recurrent and its (unique, due to its irreducibility) stationary distribution π satisfies $\sum_j \pi(j) f(j) < \infty$. Furthermore, we have $\sum_j P_{ij}^n f(j) \to \sum_j \pi_j f(j)$ as $n \to \infty$.

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press

We are now ready to state the following.

THEOREM 2: *The Markov chain* (Φ_n) *is positive recurrent with stationary probability* π_i *that has finite mean* $\sum_{i=1}^{\infty} \pi_i i < \infty$ *provided that* $E\sigma := \delta \sum_{j=1}^{\infty} j g_j < \infty$.

PROOF: We will assume that $g_j > 0$ for infinitely many j's (i.e., $\{g_j\}$ has an infinite tail); otherwise, the Markov chain (Φ_n) has finite state space and it is obviously positive recurrent. Furthermore, in this case, the mean expected stage length in steady state is also finite and thus the system is stable.

We will use the Foster–Lyapunov drift criterion of Theorem 1 with $f(i) = i$ and $V(i) := ci$, where $c > 1$. With this choice, (8) is equivalent to

$$\frac{1}{i} \sum_{j=1}^{\infty} P_{ij} j \leq 1 - c^{-1}. \tag{9}$$

Taking into account inequality (4), we see that (9) is satisfied provided that the following inequality holds:

$$\sum_{j=1}^{\infty} \beta e^{-i\beta \bar{G}_j} g_j j + \frac{1}{i} \sum_{j=1}^{\infty} e^{-i\beta} g_j j \leq 1 - c^{-1}. \tag{10}$$

By the Dominated Convergence Theorem,

$$\lim_{i\to\infty} \sum_{j=0}^{\infty} \beta e^{-i\beta \bar{G}_j} g_j j = \sum_{j=0}^{\infty} \lim_{i\to\infty} \beta e^{-i\beta \bar{G}_j} g_j j = 0, \tag{11}$$

where we have used the dominating function jg_j. (This is a valid dominating function because $\sum_{j=0}^{\infty} j g_j < \infty$. We have also used the fact that $\bar{G}_j > 0$ for all $j \in \mathbb{N}$ since, by assumption, $\{g_j\}$ has an infinite tail.) Using again the Dominated Convergence Theorem on the second term of the right-hand side of (11) with the same dominating function, we have

$$\lim_{i\to\infty} \sum_{j=0}^{\infty} \frac{e^{-i\beta}}{i} g_j j = \sum_{j=0}^{\infty} \lim_{i\to\infty} \frac{e^{-i\beta}}{i} g_j j = 0. \tag{12}$$

Clearly, (11) and (12) imply that (10) and, hence, (9) hold for all i greater than some $N \in \mathbb{N}$.

The finiteness of $\sum_j g_j j$ also immediately guarantees that (7) holds for some $b > 0$ and all $i \leq N$. Thus, appealing to Theorem 1, we conclude that the Markov chain (Φ_n) is positive recurrent when $E\sigma < \infty$ with stationary distribution π and, furthermore, that $\sum_{i=1}^{\infty} \pi_i i < \infty$. This finiteness of the first moment of the stationary distribution guarantees that the mean stage length is finite in steady state and, thus, that there exists a proper stationary regime for this system. ■

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press

Remark 1: Note that we do not claim that when $E\sigma = \infty$, (Φ_n) cannot be positive recurrent.

Thus, in view of the discussion in Section 3, we have also shown the following.

COROLLARY 1: *The condition $\sum_j g_j j < \infty$ is necessary and sufficient for the existence of a stationary distribution of the semiregenerative process (X_t).*

PROOF: The sufficiency of the moment condition follows directly from the above analysis. As for the necessity, it is enough to argue that the mean stage length is the maximum of a number of service times and, as such, stochastically greater than a service time. Thus, if $\sum_j g_j j$ diverges, the mean stage length must be infinite, which implies that a stationary regime cannot exist for the process (X_t) by standard results regarding semiregenerative processes. ■

5. NONLATTICE SERVICE TIME DISTRIBUTIONS

When the service time distribution G is not lattice, the above analysis fails. However, a similar analysis could be carried out using drift criteria for Markov chains on general state spaces (see [6]). Alternatively, one could focus on the Markov chain (K_n) of the number of customers served in each stage, which, in every case, has a denumerable state space. However, instead of adopting either of these two approaches, we will sketch a simple argument based on stochastic ordering.

If F and G are two distribution functions on $\mathbb{R}$, we say that F stochastically dominates G if $F(x) \leq G(x)$ for all $x \in \mathbb{R}$. This stochastic order (also known as strong order) will be denoted by $F \geq_{\text{st}} G$. For further background on stochastic ordering, we refer the reader to [9]. We begin with the following obvious proposition.

PROPOSITION 1: *If G and $\tilde{G}$ are two service time distributions such that $G \leq_{\text{st}} \tilde{G}$ and if (Y_n) and $(\tilde{Y}_n)$ are the corresponding service stage Markov chains, then $Y_n \leq_{\text{st}} \tilde{Y}_n$ for all $n \in \mathbb{N}$.*

PROOF: It is immediate by appropriately constructing both chains on the same probability space. ■

A distribution F on $\mathbb{R}^+$ is called lattice if there exists $\delta > 0$ such that $F\{\delta\mathbb{N}\} = 1$. Given $\delta > 0$ and a distribution function G on $\mathbb{R}^+$, we define a lattice distribution G_δ by means of

$$G_\delta(x) = \sum_{n=0}^{\infty} G(n\delta)\mathbf{1}(n\delta \leq x < (n+1)\delta).$$

Then, $G_\delta \geq_{\text{st}} G$ and, by virtue of Proposition 1, we have $Y_n^\delta \geq_{\text{st}} Y_n$ for each $n \in \mathbb{N}$ for the corresponding service stage Markov chains. Furthermore, if $\int_0^\infty xG\,dx < \infty$, then $\int_0^\infty xG_\delta\,dx < \infty$. Thus, the finiteness of the first moment of the service time distribution implies the positive recurrence of the countable state space Markov chain (Y_n^δ)

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press

and the finiteness of the first moment of *its* stationary distribution, which, in turn, implies the stability of the original system.

References

1. Asmussen, S. (1987). *Applied probability and queues*. New York: Wiley.
2. Avi-Itzhak, B. & Halfin, S. (1989). Response times in gated $M/G/1$ queues: The processor-sharing case. *Queueing Systems* 4(3): 263–279.
3. Browne, S., Coffman, E.G., Jr., Gilbert, E.N., & Wright, P.E. (1992). Gated, exhaustive, parallel service. *Probability in the Engineering and Informational Sciences* 6: 217–239.
4. Browne, S., Coffman, E.G., Jr., Gilbert, E.N., & Wright, P.E. (1992). The gated infinite-server queue: Uniform service times. *SIAM Journal on Applied Mathematics* 52(6): 1751–1762.
5. Jagerman, D.L. & Sengupta, B. (1989). A functional equation arising in a queue with a gating mechanism. *Probability in the Engineering and Informational Sciences* 3: 417–433.
6. Meyn, S.P. & Tweedie, R.L. (1993). *Markov chains and stochastic stability*. New York: Springer-Verlag.
7. Rege, K.M. & Sengupta, B. (1989). A single server queue with gated processor-sharing discipline. *Queueing Systems* 4(3): 249–261.
8. Tan, X. & Knessl, C. (1994). Heavy traffic asymptotics for a gated, infinite-server queue with uniform service times. *SIAM Journal of Applied Mathematics* 54(6): 1768–1779.
9. Stoyan, D. (1984). *Stochastic ordering*. New York: Wiley.
10. Tweedie, R.L. (2000). Markov chains: Structure and applications. In D.N. Shanbhag & C.R. Rao (eds.), *Handbook of statistics*, Vol. 19. Amsterdam: Elsevier, pp. 817–851.

https://doi.org/10.1017/S0269964804181072 Published online by Cambridge University Press