A Queue with Multiple Stable Regions

Guang-Liang Li [*] and Victor O.K. Li[†]

Abstract

The stable region of a queue consists of all values of the system parameters for the queue to be stable. A queue can have multiple stable regions, such that the probability law governing the system has different functional forms in different stable regions and hence the performance of the system cannot be captured only by the parameter values. We analyze a queue with multiple stable regions, and explain why such a queue is not amenable to current queueing theory.

1 Introduction

The following scenario illustrates the problem to be solved. A service provider operates a database and a communication link. The files in the database are retrieved and then transmitted over the link. A queueing model describes the behavior of the system. The link represents the "server". The "customers" are the retrieved files to be transmitted. The "service time" of a customer is the transmission time of the corresponding file, modeled by an exponential random variable A of rate μ with pdf f_A. The "inter-arrival time" between the nth customer and the $(n+1)$th customer for any n is a bounded random variable and no other information is available. So the inter-arrival time is modeled by a uniform random variable T with a pdf $f_T(t) = 1/h$ for $t \in (0, h)$, and $f_T(t) = 0$ otherwise [2]. The server serves the customers in a work-conserving, FCFS manner. The service times are i.i.d., distributed as the random variable A. The inter-arrival times are also i.i.d., distributed as the random variable T. The service and inter-arrival times are mutually independent. The number of buffers is sufficiently large and the system is reliable, so no customer will be blocked or lost. The mean inter-arrival time is greater than the mean service time, so the queue is stable. We call this model the U/M/1 queue. To guarantee the delay performance of the system as required by the clients, the service provider wants to know whether the asymptotic waiting time W of the queue can be bounded above.

As a consequence of implicitly assuming an unbounded number of services in an inter-arrival time during a busy period (e.g., see the infinite sum in the first equation on p. 249, and the definition of β_n on p. 245, [3]), the waiting time of any stable, ordinary G/M/1 queue with continuous service and inter-arrival times is unbounded as shown in the literature, and all the solution techniques

[*] The corresponding author. Department of Electrical and Electronic Engineering, The University of Hong Kong, `glli@eee.hku.hk`

[†] Department of Electrical and Electronic Engineering, The University of Hong Kong, `vli@eee.hku.hk`

in current queueing theory are subject to the above assumption. However, the assumption is actually incorrect due to a misunderstanding on the meaning of the stability of a queue. We discuss this issue in Section 2. With the results proved in Section 3, we can indeed find the pdf of the bounded waiting time. In Section 4, we analyze the behavior of the U/M/1 queue.

The analysis reveals an interesting phenomenon: The U/M/1 queue has multiple stable regions. As the load increases, i.e., as μh decreases, the asymptotic behavior of the queue can change qualitatively, characterized by different functional forms of the pdf of W. Accordingly, the performance of the queue, characterized by the statistics of the asymptotic waiting time, degrades, if the parameters (μ, h) vary from a stable region to another stable region. The degradation cannot be captured just by the values of the parameters, since the functional form of the waiting-time pdf has changed. Multiple stable regions do not contradict the uniqueness of the asymptotic behavior of the queue, since different asymptotic behaviors correspond to different values of the system parameters. Once the values of the parameters are given, the asymptotic behavior is uniquely determined by the values. But when the parameters vary across a critical boundary within the stable region, the asymptotic behavior changes qualitatively. This is an essential difference between the U/M/1 queue and any queue without multiple stable regions.

Although our analysis is within the framework of queueing theory, the results can be described in the language of catastrophe theory [6]. The catastrophe-theoretical idea was first used for computer performance modeling in [4]. Some other researchers also used the idea in their work (e.g., [5]). The notions "multiple stable regions" and "multistability" are different. The probability law of a system with multistability may have a unique functional form, which allows the system to have several stable operation modes. So a system with multistability may not necessarily have multiple stable regions. One may study a system with multistability based on catastrophe theory (e.g., [4]) or using other approaches (e.g., [1]).

2 Meaning of Stability

In current queueing theory, the stability of an ordinary G/G/1 queue can imply an unbounded queue size as a result of a seemingly feasible probability assignment over all nonnegative integers (representing all possible values of the queue size) under the stability condition. In other words, for any given positive integer n, by observing the queue for a sufficiently long time, one can see n customers in the system. Such claim is incorrect.

Theorem 1 *1) The asymptotic range of the size of a stable, ordinary G/G/1 queue has a maximum $M < \infty$.*

2) The maximum possible, asymptotic ranges of the waiting and sojourn times of a stable, ordinary G/G/1 queue are $[0, H]$ and $(0, G]$, respectively, where H and G are finite.

Proof: We only prove 1). The proof of 2) is similar. By the ergodic theorem, except a set Ω_0 of probability zero in the sample space Ω of the queue, each sample point in $\Omega \setminus \Omega_0$ is representative of the population and hence contains all information of the asymptotic distribution of the queue size. Consider such a sample point $\omega \in \Omega \setminus \Omega_0$. The queue size at ω takes on a positive value only when the queue is in a busy period. Denote by $y_i(\omega)$ the maximum queue size in the ith busy period, and write

$$Y_m(\omega) = \max\{y_1(\omega), y_2(\omega), \cdots, y_m(\omega)\}.$$

Clearly, $Y_1(\omega), Y_2(\omega), \cdots$ form a non-decreasing sequence $\langle Y_m(\omega)\rangle$. There are only two cases: (a) $\langle Y_m(\omega)\rangle$ is bounded above and hence has a maximum $M(\omega)$ with $Y_m(\omega) \leq M(\omega) < \infty$ for each m; (b) As m increases, $Y_m(\omega) \to \infty$. One of the two statements below is true and the other is false: (i) "case (a) corresponds to the stable behavior of the queue"; (ii) "case (b) corresponds to the stable behavior of the queue". Clearly, it is absurd for statement (ii) to be true. Since ω contains all information of the asymptotic distribution of the queue size and since the asymptotic distribution is unique, $M(\omega)$ must be the same at each point in $\Omega \setminus \Omega_0$. So we can write $M(\omega)$ as M. **Q.E.D.**

3 Asymptotic Waiting Time

Consider a stable, ordinary G/G/1 queue with continuous service and inter-arrival times. Let Ω be the sample space of the queue. The pdf of the service times is f_A. The pdf of the inter-arrival times is f_T. The waiting and sojourn times of the nth customer are W_n and Q_n, with pdfs φ_n and ψ_n, respectively. Since the queue is stable, as n tends to infinity, φ_n and ψ_n converge respectively to φ and ψ, which are the pdfs of the asymptotic waiting time W and sojourn time Q of the queue, respectively. The convergence means that $\varphi_n(w) \to \varphi(w)$ and $\psi_n(q) \to \psi(q)$ at all w and q where $\varphi(w)$ and $\psi(q)$ are continuous.

Conditioned on $Q_{n-1} = q$ for $q > 0$, the value of the pdf of W_n at w is $\varphi_n(w|q)$. Its asymptotic version is $\varphi(w|q)$. Similarly, conditioned on $W_n = w$ for $w \geq 0$, the value of the pdf of Q_n at q is $\psi_n(q|w)$, with an asymptotic version $\psi(q|w)$.

We start with the state equations of the queue. The state variables are W_n and Q_n. The equations govern the state evolution. When the system operates normally, the server serves the customer in the work-conserving way. So we can establish the state equations based on flow balance.

Let $w_1 \geq 0$ be an arbitrarily given constant. Beginning with $W_1 = w_1$, the evolution of W_n and Q_n is as follows. For $\omega \in \Omega$,

$$Q_n(\omega) = A_n(\omega) + W_n(\omega), \ n \in \{1, 2, \cdots\} \tag{1}$$

where A_n is the service time of the nth customer, and

$$W_n(\omega) = \begin{cases} -T_{n-1}(\omega) + Q_{n-1}(\omega), & T_{n-1}(\omega) < Q_{n-1}(\omega) \\ 0, & T_{n-1}(\omega) \geq Q_{n-1}(\omega), \end{cases} \quad n \in \{2, 3, \cdots\} \tag{2}$$

where T_{n-1} is the time between the $(n-1)$th arrival and the nth arrival.

To describe the evolution of the state of the queue in terms of the pdfs of the service and inter-arrival times, set $n = 1$ in (1). We have

$$\psi_1(q) = f_A(q - w_1), \ \ q > 0.$$

Similarly

$$\psi_n(q|w) = f_A(q - w),\ q > 0,\ n \in \{2, 3, \cdots\}. \tag{3}$$

From (2),

$$\varphi_n(w|q) = f_T(q - w) + \delta(w)P_T\{T \geq q\},\ w \geq 0,\ n \in \{2, 3, \cdots\} \tag{4}$$

where $\delta(w)$ is the Dirac delta function, and P_T is the probability measure associated with an inter-arrival time T.

Thus, together with the initial conditions $W_1 = w_1$ and $\psi_1(q) = f_A(q - w_1)$, the following two equations describe the evolution of the state variables.

$$\varphi_n(w) = \int_0^G \varphi_n(w|q)\psi_{n-1}(q)dq,\ w \geq 0,\ n \in \{2, 3, \cdots\} \tag{5}$$

and

$$\psi_{n-1}(q) = \int_0^H \psi_{n-1}(q|w)\varphi_{n-1}(w)dw,\ q > 0,\ n \in \{3, 4, \cdots\}. \tag{6}$$

The singularity caused by the Dirac delta function $\delta(w)$ at $w = 0$ in $\varphi_n(w|q)$ will also appear in $\varphi_n(w)$ (see (4)). Since this singularity is intrinsic, the functional form of $\varphi_n(w)$ is

$$\varphi_n(w) = \zeta_n(w) + v_n(w)\delta(w)$$

with $\zeta_n(w) \geq 0$ for $w > 0$. Due to the property of $\delta(w)$, only $v_n(0)$, rather than $v_n(w)$ for all values of w, will contribute to the integral of $v_n(w)\delta(w)$. Therefore, we can write equivalently

$$\varphi_n(w) = \zeta_n(w) + v_n(0)\delta(w),\quad w \geq 0,\quad n > 1. \tag{7}$$

Let $\eta > 0$. From (7), we see

$$\lim_{\eta \to 0} \int_0^\eta \varphi_n(w)dw = \lim_{\eta \to 0} \int_0^\eta \zeta_n(w)dw + v_n(0) \lim_{\eta \to 0} \int_0^\eta \delta(w)dw = v_n(0).$$

Write

$$v(0) \overset{\text{def}}{=} \lim_{n \to \infty} v_n(0).$$

So $v(0)$ is the asymptotic probability that an arriving customer finds the server idle. Since the queue is stable and ordinary, $v(0)$ exists and $0 < v(0) < 1$. Moreover, we have the following result.

Lemma 1 *The pdf of W is*

$$\varphi(w) = \zeta(w) + v(0)\delta(w),\quad w \geq 0 \tag{8}$$

where

$$\zeta(w) = \lim_{n \to \infty} \zeta_n(w) \geq 0,\quad w > 0.$$

Proof: This result follows directly from (7) and the stability of the queue. **Q.E.D.**

Now we derive an integral equation, and show how to determine the waiting-time pdf of the queue based on the solution of the equation. Write

$$u(w) \overset{\text{def}}{=} \frac{\zeta(w)}{v(0)}, \quad w > 0. \tag{9}$$

Theorem 2 *The function $u(w)$ is the unique solution of the following integral equation.*

$$u(w) = K(w,0) + \int_0^H K(w,x)u(x)dx, \quad w \in (0,H]. \tag{10}$$

The kernel $K(w,x)$ of (10) is given by (12) below.

Proof: From (5),

$$\varphi_n(w) = \int_0^G \varphi_n(w|y)\psi_{n-1}(y)dy.$$

By using (6), $\varphi_n(w)$ can be further expressed as

$$\begin{aligned}\varphi_n(w) &= \int_0^G \varphi_n(w|y) \int_0^H \psi_{n-1}(y|x)\varphi_{n-1}(x)dxdy \\ &= \int_0^H \left[\int_0^G \varphi_n(w|y)\psi_{n-1}(y|x)dy\right] \varphi_{n-1}(x)dx.\end{aligned}$$

The left-hand side of the above equation is a pdf, so the double integral on the right-hand side converges.

Since both the service times and the inter-arrival times are associated with common pdfs, we can drop the subscripts in $\varphi_n(w|y)$ and $\psi_{n-1}(y|x)$, and have

$$K_0(w,x) = \int_0^G \varphi(w|y)\psi(y|x)dy$$

with $\varphi(w|y)$ and $\psi(y|x)$ given by (4) and (3), respectively. So

$$\varphi_n(w) = \int_0^H K_0(w,x)\varphi_{n-1}(x)dx. \tag{11}$$

Write

$$K(w,x) = \int_{\max\{w,x\}}^G f_T(y-w)f_A(y-x)dy \tag{12}$$

and

$$J(x) = \int_x^G f_A(y-x)P_T\{T \geq y\}dy.$$

We see readily

$$K_0(w,x) = K(w,x) + J(x)\delta(w). \tag{13}$$

Insert (7) and (13) into (11) and compare both sides. We have

$$\zeta_n(w) = K(w,0)v_{n-1}(0) + \int_0^H K(w,x)\zeta_{n-1}(x)dx. \tag{14}$$

From Lemma 1, a unique function $\zeta(w) > 0$ exists, such that $\zeta_n(w)$ converges to $\zeta(w)$. In (14), divide both sides by $v(0)$, and let $n \to \infty$. Equation (10) then follows. **Q.E.D.**

The following corollary shows how to obtain the pdf of the waiting time, based on the solution of (10).

Corollary 1 *The solution of (10) determines uniquely the pdf of W with $v(0)$ and $\zeta(w)$ given below by (15) and (16), respectively.*

Proof: From (8), the pdf of W is determined by $v(0)$ and $\zeta(w)$. It is sufficient to show that $v(0)$ and $\zeta(w)$ can be expressed by $u(w)$, the unique solution of (10). From (9) and (8)

$$u(w) = \frac{\zeta(w)}{v(0)} = \frac{\varphi(w) - v(0)\delta(w)}{v(0)}.$$

Integrate both sides,

$$\int_0^H u(w)dw = \frac{\int_0^H \varphi(w)dw - v(0)\int_0^H \delta(w)dw}{v(0)} = \frac{1 - v(0)}{v(0)}.$$

Thus

$$v(0) = \frac{1}{1 + \int_0^H u(w)dw}. \tag{15}$$

Insert (15) into (9), we see

$$\zeta(w) = \frac{u(w)}{1 + \int_0^H u(x)dx}. \tag{16}$$

Due to the uniqueness of $u(w)$, the pdf of W is unique. **Q.E.D.**

We can obtain known results in current queueing theory by setting $H = G = \infty$ and solving the corresponding integral equation. As an example, consider a stable, ordinary M/M/1 queue with arrival rate λ and service rate μ. We can readily see

$$u(w) = \lambda e^{-(\mu-\lambda)w}$$

which determines, via Lemma 1, the well-known pdf of W of the M/M/1 queue. However, as shown by Theorem 1, the pdf given in current queueing theory is incorrect.

4 Multiple Stable Regions of U/M/1 Queue

By (12), the kernel of the integral equation for the U/M/1 queue is

$$K(w,x) = \int_{\max\{w,x\}}^{w+h} f_T(y-w) f_A(y-x) dy = \begin{cases} \frac{1}{h}(1 - e^{-\mu h} e^{-\mu w} e^{\mu x}), & w \le x; \\ \frac{1}{h}(1 - e^{-\mu h}) e^{-\mu w} e^{\mu x}, & w \ge x. \end{cases}$$

From (10), the integral equation of the U/M/1 queue is

$$u(w) = \frac{1 - [1 + p(H)]e^{-\mu h}}{h} e^{-\mu w} + \frac{1}{h} e^{-\mu w} \int_0^w e^{\mu x} u(x) dx + \frac{1}{h} \int_w^H u(x) dx. \quad (17)$$

Write

$$p(w) = \int_0^w e^{\mu x} u(x) dx, \ q(w) = \int_w^H u(x) dx$$

where $u(x)$ is the solution of (17), and

$$E = \frac{1 - [1 + p(H)]e^{-\mu h}}{h}.$$

Consequently we can re-write (17) as

$$u(w) = E e^{-\mu w} + \frac{1}{h} e^{-\mu w} p(w) + \frac{1}{h} q(w).$$

We then readily have

$$\frac{dp(w)}{dw} = E + \frac{1}{h} p(w) + \frac{1}{h} e^{\mu w} q(w) = -e^{\mu w} \frac{dq(w)}{dw}$$

and

$$\frac{d^2 p(w)}{dw^2} - \mu \frac{dp(w)}{dw} + \frac{\mu}{h} p(w) = -\mu E. \quad (18)$$

The differential equation (18) can be easily solved, and from the definition of $p(w)$, we have

$$u(w) = e^{-\mu w} \frac{dp(w)}{dw}.$$

With $p(w)$ available, we can determine $u(w)$, and obtain the pdf of W immediately from (8), (16), and (15). Since the functional form of the pdf of W is basically determined by $p(w)$, we only consider $p(w)$ in the following discussion.

The asymptotic behavior of the queue depends on the values of the system parameters μ and h. Note that μh characterizes the load of the queue. The set of meaningful values of the parameters is

$$\{(\mu, h) : \mu > 0, \ h > 0\}.$$

The stability condition of the U/M/1 queue is

$$\mu^{-1} < h/2. \quad (19)$$

The unstable region corresponds to a set of the values of the parameters that violate the stability condition (19).

$$\{(\mu, h) : 0 < \mu h \leq 2\}.$$

The set of all values of the parameters satisfying (19), i.e.,

$$\{(\mu, h) : 2 < \mu h < \infty\}$$

corresponds to the whole stable region of the queue, and consists of three disjoint subsets, corresponding to three different stable regions. In the following, c_1 and c_2 are unknown constants to be determined.

Stable Region 1: This region corresponds to subset

$$R_1 = \{(\mu, h) : 4 < \mu h < \infty\}.$$

Write

$$m_1 = \frac{\mu}{2}\left(1 + \sqrt{1 - \frac{4}{\mu h}}\right), \; m_2 = \frac{\mu}{2}\left(1 - \sqrt{1 - \frac{4}{\mu h}}\right).$$

We have

$$p(w) = c_1 e^{m_1 w} + c_2 e^{m_2 w} - hE.$$

Stable Region 2: A different stable region corresponds to subset

$$R_2 = \{(\mu, h) : 2 < \mu h < 4\}.$$

Write

$$a = \frac{\mu}{2}, \; b = \frac{\mu}{2}\sqrt{\frac{4}{\mu h} - 1}.$$

We have

$$p(w) = e^{aw}[c_1 \cos(bw) + c_2 \sin(bw)] - hE.$$

Stable Region 3: The stable regions 1 and 2 are separated by a hyperbola, which defines the last stable region corresponding to subset

$$R_3 = \{(\mu, h) : \mu h = 4\}.$$

We have

$$p(w) = (c_1 + c_2 w)e^{\frac{\mu}{2}w} - hE.$$

As shown by the above analysis, the U/M/1 queue has multiple stable regions. Corresponding to the set $R_1 \bigcup R_2$, the behavior of the queue is structurally stable, meaning that, if the parameters (μ, h) vary within $R_1 \bigcup R_2$, i.e., either $(\mu, h) \in R_1$ or $(\mu, h) \in R_2$, the functional form of the pdf of W will not change. On the other hand, the behavior of the queue is structurally unstable, if $(\mu, h) \in R_3$. In this case, a small perturbation of the parameters may result in a qualitative change in the functional form of the pdf. When (μ, h) cross R_3 and re-enter $R_1 \bigcup R_2$, i.e., (μ, h) vary from R_1 to R_2 or from R_2 to R_1, the functional form of the pdf also changes qualitatively. As a result, the asymptotic behavior and performance of the queue change drastically. In catastrophe theory, such phenomenon is called bifurcation. The set R_3 is called the catastrophe set or bifurcation set.

It may be interesting to study the multiple stable regions in an experimental setting. Actually, it may be necessary to determine c_1 and c_2 based on measurements. Besides c_1 and c_2, $p(H)$ in the expression of E is also unknown. So we need three constraint conditions, which may be constructed with the explicit expressions of $p(w)$ and $q(w)$. For example, corresponding to R_1,

$$p(0) = c_1 + c_2 - 1 + [1 + p(H)]e^{-\mu h} = 0$$

$$p(H) = c_1 e^{m_1 H} + c_2 e^{m_2 H} - 1 + [1 + p(H)]e^{-\mu h}$$

and

$$q(0) = \frac{c_1 m_1}{\mu - m_1}[1 - e^{-(\mu - m_1)H}] + \frac{c_2 m_2}{\mu - m_2}[1 - e^{-(\mu - m_2)H}].$$

Note

$$q(0) = \int_0^H u(x)dx = \frac{1}{v(0)} - 1.$$

We can estimate the values of H and $v(0)$ by measurements in an experimental setting such as a computer-based simulation. With the estimated values of H and $v(0)$, we may determine c_1 and c_2. Different measurement techniques or experimental settings may result in different estimated values of H. The difference is due to measurement errors and does not contradict Theorem 1. We shall report our experimental results elsewhere.

References

[1] S. Grishechkin, et al., "Multistability in queues with retransmission and its relationship with large deviations in branching processes", *Theory Probab. Appl.*, vol. 47, No. 1, pp. 139-150, 2002

[2] R. Jain, The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling, John Wiley & Sons, 1991

[3] L. Kleinrock, Queueing Systems, vol. 1, John Wiley & Sons, 1975

[4] R. Nelson, "Stochastic catastrophe theory in computer performance modeling", *J. ACM*, vol. 34, no. 3, pp. 661-685, 1987

[5] K. Sakakibara, et al., "Effect of exponential backoff scheme and retransmission cutoff on the stability of frequency-hopping slotted ALOHA systems", *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 714-722, 2003

[6] R. Thom, Structural Stability and Morphogenesis, Benjamin-Addison-Wesley, 1975