The $G/G/1$ queue

Sergey Foss

The notation $G/G/1$ queue is usually referred to a single-server queue with first-in-first-out discipline and with a general distribution of the sequences of inter-arrival and service times (which are the "driving sequences" of the system). Customers are numbered $n = 0, 1, \ldots$. We assume that customer 0 arrives to a system at time $t = 0$ and finds there an initial amount of work, so has to wait for $W_0 \geq 0$ units of time for the start of its service. Let t_n be the time between the arrivals of nth and $(n+1)$st customers and s_n the service time of the nth customer. Let W_n be the waiting time (delay) of the nth customer, i.e. time between its arrival and beginning of its service. Then, for $n \geq 0$, the sequence $\{W_n\}$ satisfies *Lindley's equations*

$$W_{n+1} = (W_n + s_n - t_n)^+ \tag{1}$$

where we use the notation $x^+ = \max(0, x)$.

We assume the two-dimensional sequence $\{(s_n, t_n)\}$ to be stationary. We consider mostly the i.i.d. sequences ($GI/GI/1$ queue, here "GI" stands for "general independent"), and then the general case in brief. The list of references contains a number of recent textbooks and surveys with the whole coverage of problems discussed in the article. The key references are [1] for Section 1 and [2, 4] for Section 2.

1 $GI/GI/1$ queue

In this Section we assume that each of the sequences $\{t_n\}$ and $\{s_n\}$ consists of i.i.d. (independent and identically distributed) random variables (r.v.'s), that these two sequences are independent and do not depend on W_0. W.l.o.g. we may assume that these sequences are extended onto negative indices $n = -1, -2, \ldots$. Let $T_0 = 0$, $T_n = \sum_{i=0}^{n-1} t_i$, $n \geq 1$ and $T_n = -\sum_{i=n}^{-1} t_i$, $n \leq -1$.

Let $a = \frac{1}{\lambda} = \mathbf{E}t_n \in (0, \infty)$ denote the interarrival mean (expectation) and $b = \mathbf{E}s_n \in (0, \infty)$ the mean service time. Then $\rho = \lambda b$ is the *traffic intensity*. We consider the case of finite means only.

For the analysis of the waiting-time distributions, the key tool is the *duality* between the maximum of a random walk and the waiting time processes. Define $\xi_n = s_n - t_n$, $\mu = \mathbf{E}\xi_n = b - a$, $S_0 = 0$, $S_n = \xi_0 + \ldots + \xi_{n-1}$, $M_n = \max_{0 \leq k \leq n} S_k$. Note that $\rho < 1$ and $\mu < 0$ are equivalent.

For the random walk $\{S_n\}$, the Strong Law of Large Numbers holds, $S_n/n \to \mu$ a.s. as $n \to \infty$. So, if $\mu < 0$, then the random variable $\nu = \max\{n \ : S_n > 0\}$ is finite a.s. and $M_n = M_\nu$ for all $n \geq \nu$. Then $M = \sup_{n \geq 0} M_n = M_\nu$ and, for any n, $\mathbf{P}(M_n \neq M) \leq \mathbf{P}(\nu > n) \to 0$, as $n \to \infty$.

Proposition 1.1. *(1) For any $n \geq 0$, r.v.'s W_n and $\max(M_n, W_0 + S_n)$ are identically distributed,*

$$W_n =_D \max(M_n, W_0 + S_n). \tag{2}$$

In particular, if $W_0 = 0$, then $W_n =_D M_n$ and probability $\mathbf{P}(W_n > x)$ increases in n, for any x.

(2) If $\rho < 1$, then a limiting steady-state waiting time W exists and coincides in distribution with that of M, $W =_D M$. Further, the distribution of W_n converges to that of W in the total variation norm, that is

$$\sup_A |\mathbf{P}(W_n \in A) - \mathbf{P}(W \in A)| \leq \mathbf{P}(\nu > n) \to 0, \quad n \to \infty. \tag{3}$$

For the steady-state random variable W and for an independent of it r.v. $\xi =_D \xi_1$, the random variables $(W+\xi)^+$ and W have the same distribution,

$$(W+\xi)^+ =_D W. \tag{4}$$

(3) If $\rho = 1$ and if $\sigma^2 = \mathbf{Var}\xi_n = \mathbf{Var}t_1 + \mathbf{Var}s_1$ is finite, then the distributions of $W_n/\sqrt{n}$ converge weakly, as $n \to \infty$, to the distribution of the absulute value of a normal random variable with mean 0 and variance σ^2.
(4) If $\rho > 1$, then $W_n/n \to \mu$ a.s., as $n \to \infty$.

1.1 Regenerative structure

Consider the case $W_0 = 0$. Let $\tau_0 = 0$, $\tau = \tau_1 = \inf\{n \geq 1 \ : \ W_n = 0\}$ and, for $k \geq 1$, $\tau_{k+1} = \inf\{n > \tau_k \ : \ W_n = 0\}$. Clearly, $\tau = \inf\{n \geq 1 \ : \ S_n \leq 0\}$.

Proposition 1.2. *Assume $W_0 = 0$. If $\rho \leq 1$, then all τ_k are finite a.s. and the random elements $(\tau_k - \tau_{k-1}, (t_i, s_i, W_i), i = \tau_{k-1}, \ldots, \tau_k - 1)$ are i.i.d. Moreover, if $\rho < 1$, then $\mathbf{E}\tau < \infty$ and, for any $\alpha > 1$, $\mathbf{E}\tau^\alpha < \infty$ if and only if $\mathbf{E}s_1^\alpha < \infty$. Further, if $\rho < 1$ and if $\mathbf{E}e^{cs_1}$ is finite for some $c > 0$, then $\mathbf{E}e^{c_1\tau}$ is finite for some $c_1 > 0$.*

We may interpret $\tau_k - \tau_{k-1}$ as the number of customers served in the kth *busy* period, the duration of the kth period is the total service time

$$B_k = \sum_{i=\tau_{k-1}}^{\tau_k - 1} s_i,$$

and this busy period is followed by the *idle* period where the server is empty during

$$I_k = \sum_{i=\tau_{k-1}}^{\tau_k - 1} t_i - \sum_{i=\tau_{k-1}}^{\tau_k - 1} s_i = \sum_{i=\tau_{k-1}}^{\tau_k - 1} (t_i - s_i)$$

units of time. The sum of a busy period and of the following idle period is a *busy cycle*, $C_k = B_k + I_k = \sum_{i=\tau_{k-1}}^{\tau_k - 1} t_i$. In particular, the first idle period, I, is equal to $I = I_1 = -S_\tau$.

Proposition 1.3. *Assume $W_0 = 0$. If $\rho < 1$, then $\mathbf{E}I$ is also finite and, for any function $f : [0, \infty) \to [0, \infty)$,*

$$\mathbf{E}f((W+\xi)^-) = \frac{\mathbf{E}f(I)}{\mathbf{E}\tau} = -\mathbf{E}\xi\frac{\mathbf{E}f(I)}{\mathbf{E}I} \tag{5}$$

and, in particular,

$$\mathbf{E}(W+\xi)^- = -\mathbf{E}\xi. \tag{6}$$

Here we use notation $x^- = -\min(x, 0)$.

1.2 Moments of the stationary waiting time

Proposition 1.4. *Assume $\rho < 1$. Let W be the stationary waiting time. For any $\alpha > 0$, if $\mathbf{E}s_1^{\alpha+1} < \infty$, then $\mathbf{E}W^\alpha < \infty$. Conversely, if $\mathbf{E}W^\alpha < \infty$ and $\mathbf{E}t_1 < \infty$, then $\mathbf{E}s_1^{\alpha+1} < \infty$. Further, if, for some $k = 1, 2, \ldots$, both $\mathbf{E}s_1^{k+1}$ and $\mathbf{E}t_1^{k+1}$ are finite, then*

$$\sum_{l=0}^{k} C_{k+1}^l \mathbf{E}W^l \mathbf{E}\xi^{k+1-l} = \mathbf{E}[-(W+\xi)^{k+1}] = (-1)^k \mathbf{E}\xi \frac{\mathbf{E}I^{k+1}}{\mathbf{E}I}. \tag{7}$$

In particular,

$$\mathbf{E}W = \frac{\lambda^2(\mathbf{Var}t_1 + \mathbf{Var}s_1) + (1-\rho)^2}{2\lambda(1-\rho)} - \frac{\mathbf{E}I^2}{2\mathbf{E}I} \tag{8}$$

and then

$$\frac{\rho^2+\lambda^2\mathbf{Var}s_1-2\rho}{2\lambda(1-\rho)}\le \mathbf{E}W\le \frac{\lambda(\mathbf{Var}t_1+\mathbf{Var}s_1)}{2(1-\rho)}. \tag{9}$$

Here again random variables W and $\xi =_D \xi_1$ are assumed to be independent.

1.3 Continuous time. The virtual waiting time

Recall that $C=C_1$, $B=B_1$ and $I=I_1$ are, correspondingly, the (duration of) the first busy cycle, the first busy period, and the first idle period.

Proposition 1.5. *Assume $W_0=0$. Suppose $\rho<1$. Then the busy cycle has mean $\mathbf{E}C=a\mathbf{E}\tau$, the busy period has mean $\mathbf{E}B=b\mathbf{E}\tau$, and the mean of the idle period is $\mathbf{E}I=\mathbf{E}C-\mathbf{E}B=-\mu\mathbf{E}\tau$.*

Let $W(t)$ be the *virtual waiting time* at time t, i.e. the total residual workload at t which is the sum of the residual service time of a customer in service and of all service times of customers in the queue. As a function of t, $W(t)$ decreases linearly between consequtive arrivals if positive and stays at zero if zero, with positive jumps at the arrival epochs. Further, $W_n=W(T_n-)$, $n=0,1,\dots$ and the process $W(t), t\ge 0$ satisfied *Lindley's equations in continuous time*:

$$W(t)=(W(T_n-)+s_n-(t-T_n))^+,\quad t\in[T_n,T_{n+1}) \tag{10}$$

Non-Lattice distribution. A random variable X has a *non-lattice* distribution if $\sum_{k=-\infty}^{\infty}\mathbf{P}(X=kh)<1$, for all $h>0$.

Proposition 1.6. *Suppose that $\rho<1$ and that $\{t_n\}$ have a common non-lattice distribution. Then there exists a limiting (stationary) distribution, as $t\to\infty$, of the virtual time $W(t)$ which is given by*

$$\mathbf{P}(V>x)=\frac{1}{\mathbf{E}C}\mathbf{E}\left(\int_0^C \mathbf{I}(W(u)>x)du\right). \tag{11}$$

Further,

$$\mathbf{E}V=\rho\left(\frac{\mathbf{E}s^2}{2\mu_s}+\mathbf{E}W\right). \tag{12}$$

Here $\mathbf{I}(\cdot)$ is the indicator function, $\mathbf{I}(A)=1$ if event A occurs and $\mathbf{I}(A)=0$, otherwise.

Proposition 1.7. *Assume that a random variable $\widehat{s}$ is independent of W and has an absolutely continuous distribution with density $\mathbf{P}(s>x)/b$ where $b=\mathbf{E}s_1$. Then*

$$\mathbf{P}(V=0)=1-\rho$$

and, for all $x\ge 0$,

$$\mathbf{P}(V>x)=\rho\mathbf{P}(W+\widehat{s}>x).$$

1.4 Queue Length and Little's law

We introduce three quantities: the queue length Q_n^A at the n arrival time, the queue length Q_n^D at the n departure time, and the queue length $Q(t)$ at an abritrary time t.

Spread-out distribution. A probability distribution F is *spread-out* if it can be represented as $F=pG_1+(1-p)G_2$ where $p\in[0,1)$, G_1 and G_2 are probability distributions, and G_2 is abolutely continuous with respect to Lebesgue measure.

Proposition 1.8. *If $\rho < 1$, then the distribution of Q_n^A (correspondingly, of Q_n^D) converges, as $n \to \infty$, to that of a finite random variable Q^A (correspondingly, Q^D), in the total variation norm.*
If $\rho < 1$ and the distribution of the inter-arrival times is non-lattice, then the distribution of $Q(t)$ converges weakly, as $t \to \infty$, to that of a finite random variable Q. If, in addition, the common distribution of inter-arrival times is spread-out, then the convergence is in the total variation norm.

Proposition 1.9. *(*Little's law*) Assume that $\rho < 1$, $\mathbf{E}s_1^2 < \infty$, and that the distribution of inter-arrival times is non-lattice. Then*

$$q = \lambda(w + b) \tag{13}$$

where $q = \mathbf{E}Q$ is the mean stationary queue length, $\lambda = 1/a$ the arrival rate, and $w + b = \mathbf{E}(W + s_1)$ the mean stationary sojourn time, i.e. the sum of mean waiting and mean service times.

Proposition 1.10. *(*Distributional Little's law in discrete time*) Suppose $\rho < 1$. Then, for any $k = 1, 2, \ldots$,*

$$\mathbf{P}(Q^A \geq k) = \mathbf{P}(Q^D \geq k) = \mathbf{P}(W + s_1 \geq \sum_{i=1}^{k} t_i) \tag{14}$$

where the random variables W, s_1 and $\{t_i\}_{i=1}^k$ are assumed to be mutually independent. Further, if either s_1 or t_1 has a continuous distribution, then equations (14) are equivalent to

$$\mathbf{P}(Q^A = 0) = \mathbf{P}(Q^D = 0) = \mathbf{P}(W = 0) \tag{15}$$

and, for $k \geq 1$,

$$\mathbf{P}(Q^A \geq k) = \mathbf{P}(Q^D \geq k) = \mathbf{P}(W > \sum_{i=1}^{k-1} t_i). \tag{16}$$

Proposition 1.11. *(*Distributional Little's law in continuous time*) Suppose that $\rho < 1$ and that the distribution of inter-arrival times is non-lattice. Then the stationary distribution of the queue length Q is given by*

$$\mathbf{P}(Q = 0) = 1 - \rho$$

and, for $k = 1, 2, \ldots$,

$$\mathbf{P}(Q \geq k) = \mathbf{P}(V > \sum_{i=1}^{k-1} t_i) = \rho \mathbf{P}(W + \widehat{s} > \sum_{i=1}^{k-1} t_i) \tag{17}$$

where we assume the mutual independence of all r.v.'s in the formula above.

1.5 Continuity under perturbation

Proposition 1.12. *Consider a series of single-server queue indexed by the upper k. Assume that, as $k \to \infty$, the distributions of $t_1^{(k)}$ weakly converge to that of t_1, the distributions of $s_1^{(k)}$ weakly converge to that of s_1, and that $\rho < 1$ and $\rho^{(k)} < 1$, for all k. If $\mathbf{E}s_1^{(k)} \to \mathbf{E}s_1$, then the distributions of $W^{(k)}$ converge weakly to that of W. If, in addition, either the distribution of s_1 or the distribution of t_1 is continuous, then $\mathbf{P}(W^{(k)} = 0) \to \mathbf{P}(W = 0)$.*

1.6 Heavy-traffic limits

Proposition 1.13. *Consider a series of single-server queues indexed by the upper index k. Assume that, as $k \to \infty$, the distributions of $t_1^{(k)}$ weakly converge to that of t_1, the distributions of $s_1^{(k)}$ weakly converge to that of s_1, and that the distributions of s_1 and of t_1 do not both degenerate. Assume further that $\rho^{(k)} < 1$, for all k, and $\rho^{(k)} \to \rho = 1$, and that the squares $\left(s_1^{(k)}\right)^2$ and $\left(t_1^{(k)}\right)^2$ are uniformly integrable.*
Then the distributions of random variables $Y^{(k)} = |\mu^{(k)}| W^{(k)} \left(\sigma^{(k)}\right)^2$ converge weakly to an exponential distribution with intensity 2 and also $\mathbf{E}Y^{(k)} \to \frac{1}{2}$. Furthermore, for each T and for $y \geq 0$,

$$\mathbf{P}\left(\frac{|\mu^{(k)}|}{\left(\sigma^{(k)}\right)^2} W^{(k)}_{[T\left(\sigma^{(k)}\right)^2]} > y \right) \to \mathbf{P}\left(\max_{0 \leq t \leq T} (B(t) - t) \geq y \right) \tag{18}$$

where $\{B(t), t \geq 0\}$ is the standard Brownian motion, and $[x]$ is the integer part of a number x.

1.7 Rare events

In the stable case, $\rho < 1$, the exact values of the tail probability $\mathbf{P}(W > x)$ may be found only in a number of particular cases. But the asymptotics for this probability may be written down in a great generality. There are two particular cases, of light tails and of heavy tails of the service time distributions, where the tail asymptotics differ and are caused by different phenomena.

For two positive functions f and g on the real line, we write $f(x) \sim g(x)$ if $\lim \frac{f(x)}{g(x)} = 1$ as $x \to \infty$.

1.7.1 Rare events in the presence of light tails

Proposition 1.14. *Assume that $\rho < 1$ and the distribution of the service times has the light tail, $\mathbf{E}e^{ds_1} < \infty$ for some $d > 0$ and, moreover, that there exists a constant $\gamma > 0$ such that $\mathbf{E}e^{\gamma(s_1 - t_1)} = 1$ and $d := \mathbf{E}(s_1 - t_1)e^{\gamma(s_1 - t_1)} \in (0, \infty)$. Assume further that the distribution of r.v. $s_1 - t_1$ is non-lattice. Then, as $x \to \infty$,*

$$\mathbf{P}(W > x) \sim re^{-\gamma x} \tag{19}$$

where $r = \mathbf{E}e^{-\gamma\chi} \in (0, 1)$ and χ is the overshoot over the infinite barrier (see, e.g., [1] or [2] for the definition and more detail). Further, assume that $W = W_0$ is the stationary waiting time of customer 0 in the system that runs, say, from time $-\infty$. As $x \to \infty$,

$$\mathbf{P}(A(x) \mid W_0 > x) \to 1 \tag{20}$$

where $A(x)$ is the event of the form:

$$A(x) = \{ \sum_{i=-[x/d]}^{-[x/d]+j-1} (s_i - t_i) \in (-R(x) + j(d - \varepsilon(x)), R(x) + j(d + \varepsilon(x)), \quad \text{for} \quad j = 1, 2, \ldots, [x/d]\}$$

and $R(x)$ is any function tending to infinity and $\varepsilon(x)$ is any function tending to zero, as $x \to \infty$. Roughly speaking, the latter means that, given the event $\{W_0 > x\}$ occurs, all increments $(s_i - t_i)$, are approximately equal to d, for $i = -[x/d], -[x/d] + 1, \ldots, -1$.

In particular, the distribution of $s_1 - t_1$ is non-lattice if either of the distributions of s_1 or t_1 is non-lattice.

1.7.2 Rare events in the presence of heavy tails

Here we assume that the distribution of service times is *heavy-tailed*, i.e. does not have finite exponential moments, $\mathbf{E}e^{\alpha s_1} = \infty$, for all $\alpha > 0$.

A distribution F on the positive half-line with an unbounded support is *subexponential* if

$$\overline{F * F}(x) \sim 2\overline{F}(x), \quad \text{as} \quad x \to \infty.$$

It is known that any subexponential distribution is *long-tailed*, i.e. $\overline{F}(x+c) \sim \overline{F}(x)$ as $x \to \infty$, for any constant c. Further, any long-tailed distribution is heavy-tailed.

Typical examples of subexponential distributions are log-normal distributions, distributions with power tails $(1+x)^{-\alpha}$, $\alpha > 0$ and Weibull tails e^{-x^β}, $\beta \in (0,1)$.

Let again r.v. $\widehat{s}$ have the absolutely continuous distribution with distribution function $F_{\widehat{s}}(x)$ and density $f_{\widehat{s}}(x) = \mathbf{P}(s > x)/b$. Clearly, the distribution of service time s is heavy-tailed if and only if the distribution of $\widehat{s}$ is.

Proposition 1.15. *Assume that $\rho < 1$ and that the distribution of $\widehat{s}$ is subexponential. Then, as $x \to \infty$,*

$$\mathbf{P}(W > x) \sim \frac{\rho}{1-\rho}(1 - F_{\widehat{s}}(x)) = \frac{1}{a-b}\int_x^\infty \mathbf{P}(s_1 > y)\, dy. \tag{21}$$

Further, assume that $W = W_0$ is the stationary waiting time of customer 0 in the system that runs, say, from time $-\infty$. As $x \to \infty$,

$$\mathbf{P}(B(x) \mid W_0 > x) \to 1 \tag{22}$$

where $B(x)$ is the event of the form:

$$B(x) = \bigcup_{i \geq 1} \{(s_{-i} > x + i(a-b)\}.$$

This means that, for large x, the main cause for the event $\{W_0 > x\}$ to occur is a single big jump of one of previous service times.

2 General $G/G/1$ queue

In this Subsection we consider the single-server queue under the the stationary ergodic assuptions on the driving sequences. Namely, the sequence $\{t_n, s_n)\}$ is *stationary* if the joint distribution of random vectors $\{(t_{m+i}, s_{m+i})\}_{0 \leq i \leq k}$ does not depend on m, for any k. In addition, this sequence is *ergodic* if, for any event A generated by the driving sequence, the equality $\mathbf{P}(A \bigcap A^\theta) = \mathbf{P}(A)$ implies that $\mathbf{P}(A)$ is either 0 or 1. Here A^θ is defined as follows. Any event A generated by the driving sequence may be represented as $A = \{g(t_i, s_i; i = \ldots, -1, 0, 1, \ldots) = 1\}$ where g is a measurable function of the driving sequence. Then $A^\theta = \{g(t_{i+1}, s_{i+1}; i = \ldots, -1, 0, 1, \ldots) = 1\}$. In particular, the sequence is ergodic if it is i.i.d. or, more generally, the tail sigma-algebra generated by this sequence is *trivial*, i.e. contains only events of probability 0 or 1.

Proposition 2.1. *Under the stability condition $\rho = \mathbf{E}s_1/\mathbf{E}t_1 < 1$, there exists a unique stationary workload process $\widetilde{W}(t), -\infty < t < \infty$ which satisfies equations (10) on the whole real line and is such that*

$$\widetilde{W}(0) = \sup_{n \leq 0}\left(T_n + \sum_{i=0}^n s_i\right)^+. \tag{23}$$

Further, there is an infinite number of negative indicies n and infinite number of positive indices n such that $\widetilde{W}(T_n -) = 0$.

If $\rho > 1$, then there is no a finite staionary workload process and, for any initial condition $W_0 \geq 0$, $W(t)/t \to \rho - 1$ a.s., as $t \to \infty$.
If $\rho = 1$, then there may or may not exist a finite stationary process.

[**Comment to the Editors:** I discuss only FCFS $G/G/1$ queue and DO NOT consider a number of related topics, like batch arrivals, other service disciplines, etc. Hope they are covered in other articles.]

3 References

1. S. Asmussen (2003) *Appried Probability and Queues*, 2nd Edition. Springer-Verlag.
2. F. Baccelli and P. Bremaud (2003) *Elements of Queueing Theory*, 2nd Edition. Springer-Verlag.
3. A.A. Borovkov (1976) *Stochastic Processes in Queueing Theory*. Springer-Verlag.
4. A. Brandt, P. Franken and B. Lisek (1992) *Stationary Stochastic Models*. Wiley.
5. P. Bremaud (1999) *Markov Chains, Gibbs Fields, Monte-Carlo Simulation and Queues*. Springer-Verlag.
6. J.W. Cohen (1982) *The Single Server Queue*. North-Holland.
7. D.R. Cox and W.L. Smith (1961) *Queues*. Methuen.
8. B.V. Gnedenko and I.N. Kovalenko (1989) *An Introduction to Queueing Theory*, 2nd Edition. Burkhauser.
9. V.V. Kalashnikov (1994) *Mathematical Methods in Queueing Theory*. Kluwer.
10. L. Kleinrock (1975, 1976) *Queueing Systems* **I, II**. Wiley.
11. M. Miyazawa (1994) *Rate Conservation Laws: A Survey*. Queueing Systems Theory Appl., **15**, 1–58.
12. Ph. Robert (2003) *Stochastic Networks and Queues*. Springer-Verlag.
13. W. Whitt (2002) *Stochastic-Process Limits*. Springer-Verlag.