Mathematical Problems in Engineering
Volume 6, pp. 1–11
Reprints available directly from the publisher
Photocopying permitted by license only

The Queue-Length in GI/G/s Queues

PIERRE LE GALL

France Telecom, CNET, 4 Parc de la Berengere,
F-92210 Saint-Cloud, France

(Received 21 October 1998; In final form July 1999)

The distribution of the queue-length in the stationary symmetrical GI/G/s queue is given with an application to the M/G/s queue, particularly in the case of the combination of several packet traffics, with various constant service times, to dimension the buffer capacity.

Keywords: GI/G/s queue; M/G/s queue; Queue-length

1. INTRODUCTION

In a recent paper [1], we studied and evaluated the queueing delay in the stationary G/G/s queue, and particularly in the GI/G/s queue. In this paper we propose to evaluate the queue length in the stationary symmetrical GI/G/s queue only, to avoid the impact of some possible dependence between the arrival process and the queueing process.

From a practical point of view, we will consider the symmetrical M/G/s queue, with a numerical application to the packet traffic with various constant packet lengths (i.e. various constant service times), to dimension the buffer capacity.

2. NOTATION, ASSUMPTIONS AND PRELIMINARY RESULTS FOR THE SYMMETRICAL GI/G/s QUEUE

2.1. Notation and Assumptions

2.1.1. The Arrival Process

The arrival process is a *renewal process* of successive arrivals X_n, the successive *arrival intervals* Y_n being *mutually independent* and *identically distributed*, with $F_0(t)$ for distribution function. We let, for $\mathrm{Re}(z) < 0$:

$$\varphi_0(z) = E[\mathrm{e}^{zY_n}], \quad \text{and} \quad \alpha_1(z) = \int_0^\infty \mathrm{e}^{zt} \cdot \rho(t) \cdot \mathrm{d}t = \frac{\varphi_0(z)}{1 - \varphi_0(z)}, \quad (1)$$

where $\rho(t)$ is the arrival rate at time $(t + t_0)$ if an arbitrary arrival occurred at time t_0. We assume $\varphi_0(z)$ to be *holomorphic* at the origin. From Paul Levy's theorem, we deduce that $\varphi_0(z)$ *exists for* $\mathrm{Re}(z) < \delta$ *where* δ *is a positive real number*. The stationary assumption and the Abelian theorem give that $\lim_{z \to 0} z \cdot \alpha_1(z) = \Lambda$, where Λ *is the mean total arrival rate*.

2.1.2. The Service Times

The successive service times T_n are *mutually independent* and independent of the arrival process. The service times are *identically distributed* with a distribution function $F_1(t)$ and we let: $\varphi_1(z) = E[\mathrm{e}^{zT_n}]$, for $\mathrm{Re}(z) < 0$. As for $\varphi_0(z)$, we assume $\varphi_1(z)$ to be *holomorphic* at the origin and, consequently, we deduce the existence of $\varphi_1(z)$ for $\mathrm{Re}(z) < \delta$ $(\delta > 0)$.

2.1.3. The Service Discipline

The servers are *indistinguishable* for the service discipline which is "*first come-first served*". In fact, we will use this assumption for clarity in the reasoning but we know that there is no impact on the results in the queue-length.

2.1.4. The Traffic Handled

The *traffic intensity* (per server) is denoted by $\rho = (\Lambda/s) \cdot T = \lambda \cdot T$ ($\rho < 1$), where T is the *mean service time* and λ is the *mean arrival rate per server.*

2.1.5. Queueing Delay

Since the term "*waiting time*" means "*sojourn time*" in Little's formula, for clarity we prefer to use the term "*queueing delay*" for the queueing process only.

2.1.6. Contour Integrals

In this paper we use (Cauchy) contour integrals along the imaginary axis in the complex plane. If the contour (followed from the bottom to the top) is to the right of the imaginary axis (the contour being closed at infinity to the right), we write $\int_{+0}$. If the contour is to the left of the imaginary axis, we write $\int_{-0}$. Unless it is necessary to specify whether the contour is to the right or to the left of the imaginary axis, we write $\int_0$.

2.2. Preliminary Results for the Symmetrical GI/G/s Queue

In [1] we presented the following results for the *symmetrical* GI/G/s queue (= multiserver queue), in a *stationary regime*, except for the GI/D/s queue. This latter case has to be excluded because of a deterministic mechanism for the choice of arrivals.

2.2.1. Behavior of Each Server and Delayed Customer

Each server behaves as a GI/G/1 server, as if an arbitrary arrival is chosen with probability $(1/s)$ of being handled by the considered server. Consequently, in [1] the arrival rate $\rho(t)$ becomes $[\rho(t)/s]$ per server, and expression (1) becomes for each server:

$$\alpha_1(z) \to \frac{1}{s} \cdot \frac{\varphi_0(z)}{1 - \varphi_0(z)} = \frac{\alpha_1(z)}{s}. \tag{2}$$

If w_0 is the queueing delay due to this server, Pollaczek's formula [2] gives, for $\mathrm{Re}(q) \geq 0$:

$$E[\mathrm{e}^{-qw_0}] = \exp\left\{\frac{-1}{2\pi i} \cdot \int_{+0} \left[\frac{1}{q+\zeta} - \frac{1}{\zeta}\right] \cdot \log K(\zeta) \cdot \mathrm{d}\zeta\right\},$$

$$\text{with } K(\zeta) = 1 - \frac{\alpha_1(-\zeta)}{s} \cdot [\varphi_1(\zeta) - 1]. \tag{3}$$

We may note that

$$\lim_{z \to 0}\left(z \cdot \frac{\alpha_1(z)}{s}\right) = \frac{\Lambda}{s} = \lambda, \tag{4}$$

where λ is the *mean arrival rate per server*. When $|q|$ increases indefinitely, we obtain the following expression giving the probability of no delay of the considered server:

$$Q_0 = \exp\left\{\frac{1}{2\pi i} \int_{+0} \log K(\zeta) \cdot \frac{\mathrm{d}\zeta}{\zeta}\right\} = \frac{1}{n}, \tag{5}$$

n being the *mean busy period size* (= mean number of customers served) of the considered server. If P_0 $(= 1 - Q_0)$ is the probability of delay of this server, the distribution function of the queueing delay w_0 may be deduced from expression (3) and may be written as:

$$W_0(t) = 1 - P_0 \cdot G_0(t), \tag{6}$$

where $G_0(t)$ is the *complementary* distribution function of the *delayed customer*.

Finally, the *complementary* distribution function of the queueing delay due to the *multiserver* is, for a *delayed* customer,

$$G(t) = [G_0(t)]^s. \tag{7}$$

We deduce the distribution function of the *queueing delay due to the multiserver*, for an arbitrary (delayed or not) customer:

$$W(t) = 1 - P \cdot G(t) = 1 - P \cdot [G_0(t)]^s, \tag{8}$$

where P is the probability of delay *due to the multiserver*.

2.2.2. Probability of Delay

During the busy period of the multiserver (= congestion state) server's behavior has been defined in a quite independent way of partial occupancy states. For these states, it follows that a busy period (per server) appears exactly as a unique congestion state in the *lost call model*, handling n successive service times as if it was for a unique arrival but to be evaluated as n arrivals congested, n being the *mean busy period size per server*, as defined by expression (5). Consequently, the *probability of delay due to the multiserver* and used in expression (8) is:

$$P = \frac{n \cdot P_a}{1 + (n-1) \cdot P_a}, \tag{9}$$

where P_a is the *probability of loss* in the lost call model.

2.2.3. The M/G/s Queue

For the M/G/s queue, we have Poisson arrivals with expression (2):

$$\alpha_1(-z) = \frac{1}{s} \cdot \frac{\Lambda}{z} = \frac{\lambda}{z}. \tag{10}$$

Each server behaves as an M/G/1 server with Poisson arrivals, the mean arrival rate being λ. Expressions (3) and (6) become:

$$\begin{aligned} E[\mathrm{e}^{-qw_0}] &= (1-\rho) \cdot \frac{q}{q - \lambda + \lambda \cdot \varphi_1(-q)}, \\ W_0(t) &= \frac{1-\rho}{2\pi i} \cdot \int_{+0} \mathrm{e}^{qt} \cdot \frac{q}{q - \lambda + \lambda \cdot \varphi_1(-q)} \cdot \frac{\mathrm{d}q}{q}. \end{aligned} \tag{11}$$

(a) *First expression of $W_0(t)$, for each server*
On the contour (to the right of the imaginary axis), we may write

$$\begin{aligned} E[\mathrm{e}^{-qw_0}] &= \frac{1-\rho}{1 - \lambda T \cdot \varphi_1^*(-q)} = (1-\rho) \cdot \sum_{\nu=0}^{\infty} \rho^{\nu} \cdot (\varphi_1^*(-q))^{\nu}, \\ \text{with } \varphi_1^*(-q) &= \frac{1 - \varphi_1(-q)}{T \cdot q}, \end{aligned} \tag{12}$$

$\varphi_1^*(z)$ corresponding to the *remaining service time* $\tilde{T}_0$. We deduce for expression (6):

$$W_0(t) = 1 - (1-\rho) \cdot \sum_{\nu=1}^{\infty} \rho^{\nu} \cdot \left[1 - \left(\int_0^t \frac{1 - F_1(u)}{T} \cdot \mathrm{d}u \right)^{(\nu)} \right], \qquad (13)$$

where $(\cdot)^{(\nu)}$ denotes the ν-fold convolution corresponding to $[\varphi_1^*(z)]^{\nu}$.

(b) *Second expression of $W_0(t)$, for each server*
Another very useful expression of $W_0(t)$ has been given by Prabhu [4], p. 90, formula (2.106):

$$W_0(t) = 1 - (1-\rho) \cdot \sum_{i=0}^{\infty} \int_{u=0}^{\infty} \mathrm{e}^{-\lambda u} \cdot \frac{(\lambda u)^i}{i!} \cdot \mathrm{d}_u F_i(t+u), \qquad (14)$$

where $F_i(t)$ denotes the i-fold convolution of $F_1(t)$ and corresponds to $[\varphi_1(z)]^i$.

(c) *Asymptotic expression of $W_0(t)$, for each server*
The asymptotic expression of $W_0(t)$, for t large, corresponds to the (real) singularity of $E[\mathrm{e}^{-qw_0}]$ closest to the origin: $q = -\beta_0$ $(\beta_0 > 0)$. This expression is given in [2], p. 27:

$$W_0(t) \cong W_1(t) = 1 - \frac{1-\rho}{\lambda \cdot \varphi_1'(\beta_0) - 1} \cdot \mathrm{e}^{-\beta_0 t}, \qquad (15)$$

where β_0 is the real (positive) root closest to the origin of the equation

$$q + \lambda - \lambda \cdot \varphi_1(q) = 0. \qquad (16)$$

This practical expression (15) may be used for approximately $t > 5 \cdot T_0$ (T_0: mean remaining service time) and $\rho < 0.8$.

(d) *Probability of delay P of the multiserver*
For the M/G/s queue, expression (5) gives for the mean busy period size per server:

$$n = \frac{1}{1-\rho}. \qquad (17)$$

Consequently, expression (9) of P becomes the delay Erlang formula:

$$E_{2,s}(\Lambda \cdot T) = E_{2,s}(s \cdot \rho). \tag{18}$$

(e) *Expression of $W(t)$ for the multiserver*
Finally, expression (8) may be written as:

$$W(t) = 1 - P \cdot \left[\frac{1 - W_0(t)}{\rho}\right]^s, \tag{19}$$

where P is the *delay Erlang formula* $E_{2,s}(s \cdot \rho)$, $W_0(t)$ corresponding to each server. We recall that we exclude the queue M/D/s, Eq. (19) however being a good approximation.

3. THE QUEUE-LENGTH IN THE GI/G/s QUEUE

For the nth arrival in the GI/G/s queue, we denote by X_n and τ_n the arrival epoch and the queueing delay, respectively. In [3], p. 29, Pollaczek gave the condition to find j customers waiting at the epoch X_{n+j}:

$$X_{n-1} + \tau_{n-1} < X_{n+j} < X_n + \tau_n. \tag{20}$$

It follows that the condition to find at least j customers is:

$$\tau_n - (X_{n+j} - X_n) = \tau_n - (Y_1 + \cdots + Y_j) > 0. \tag{21}$$

Pollaczek worked in the complex plane to define the singular points. Now, with the facilities given by the electronic computers for the numerical calculations, we will work in the real plane. Due to (21), if $(Y_1 + \cdots + Y_j) = t$, we must have $\tau_n > t$. Consequently, in the stationary regime, the probability $P(\geq j)$ *to find at least j customers waiting just before the arrival* X_n is

$$P(\geq j) = \int_0^\infty [1 - W(t)] \cdot \mathrm{d}[F_0(t)]^{(j)}, \tag{22}$$

where $[F_0(t)]^{(j)}$ denotes the j-fold convolution of $F_0(t)$, which is the distribution function of any arrival interval. From the relation

$$\sum_{j=1}^{\infty} \mathrm{d}[F_0(t)]^{(j)} = \Lambda\, \mathrm{d}t,$$

we may deduce the *mean queue-length*:

$$L = \sum_{j=1}^{\infty} P(\geq j) = \Lambda \cdot \int_0^{\infty} [1 - W(t)] \cdot \mathrm{d}t = \Lambda \cdot \bar{W}. \qquad (23)$$

It is Little's formula applied to waiting customers only. To use simple expressions of a single server, expressions (6), (7) and (8) give

$$P(\geq j) = P \cdot \int_0^{\infty} \left[\frac{1 - W_0(t)}{P_0}\right]^s \cdot \mathrm{d}[F_0(t)]^{(j)}. \qquad (24)$$

In expression (23), the *mean queueing delay of the multiserver* becomes

$$\bar{W} = P \cdot \int_0^{\infty} \left[\frac{1 - W_0(t)}{P_0}\right]^s \cdot \mathrm{d}t. \qquad (25)$$

For numerical calculations we will consider the M/G/s queue, especially.

4. THE QUEUE-LENGTH IN THE M/G/s QUEUE

4.1. The General Expression

In the case of a Poisson arrival process we may write

$$\mathrm{d}[F_0(t)]^{(j)} = H(t,j-1) \cdot \Lambda\, \mathrm{d}t, \quad \text{with } H(t,j) = \mathrm{e}^{-\Lambda t} \cdot \frac{(\Lambda t)^j}{j!}. \qquad (26)$$

Expression (24) becomes, referring to (18),

$$P(\geq j) = E_{2,s}(s \cdot \rho) \cdot \int_0^{\infty} \left[\frac{1 - W_0(t)}{\rho}\right]^s \cdot H(t,j-1) \cdot \Lambda\, \mathrm{d}t, \qquad (27)$$

where $W_0(t)$ is defined by expression (14). Frequently, the calculation of $[1 - W_0(t)]$ is long. It is better to use the asymptotic expression as far as possible. For $\rho < 0.8$, it is appropriate to split up the preceding integral into two parts:

$$E_{2,\mathrm{s}}(s \cdot \rho) \cdot \int_0^{\infty} \left[\frac{1 - W_0(t)}{\rho}\right]^s \cdot H(t,j-1) \cdot \Lambda\,\mathrm{d}t \cong I_1(j) + I_2(j), \quad (28)$$

with

$$I_1(j) = E_{2,s}(s \cdot \rho) \cdot \int_0^{t_0} \left[\frac{1 - W_0(t)}{\rho}\right]^s \cdot H(t,j-1) \cdot \Lambda\,\mathrm{d}t,$$

$$I_2(j) = E_{2,s}(s \cdot \rho) \cdot \int_{t_0}^{\infty} \left[\frac{1 - W_1(t)}{\rho}\right]^s \cdot H(t,j-1) \cdot \Lambda\,\mathrm{d}t,$$

where $W_1(t)$ is *asymptotic* expression (15) for each server. t_0 may be approximately equal to $[5 \cdot T_0/s]$, T_0 being the *mean remaining service time*. Practically, for $P(\geq j) \leq 10^{-3}$ the term $I_1(j)$ *may be neglected* (if $\rho < 0.8$).

4.2. Case of Packet Traffics

As an important example, consider the case of a total traffic stream with N component partial Poisson traffic streams labelled j $(j = 1, \ldots, N)$. For traffic stream j, packet lengths are constant (deterministic) and equal to T_j $(T_1 < T_2 < \cdots < T_N)$. The arrival rate is Λ_j and the total arrival rate is $\Lambda = \sum_{i=1}^{N} \Lambda_j$. *Per server*, these arrival rates become $\lambda_j = [\Lambda_j/s]$ and $\lambda = [\Lambda/s]$, respectively, and the loads (traffic intensities) become $\rho_j = \lambda_j \cdot T_j$, and $\rho = \sum_{j=1}^{N} \rho_j$. For the service times, $\varphi_1(z)$ becomes

$$\varphi_1(z) = \sum_{j=1}^{N} \frac{\lambda_i}{\lambda} \cdot \mathrm{e}^{T_j z}. \quad (29)$$

In expression (14), $F_i(t)$ corresponds to

$$[\varphi_1(z)]^i = \sum_{\substack{j_1, \ldots, j_N \\ (j_1 + \cdots + j_N = i)}} \frac{i!}{j_1! \cdots j_N!} \cdot \left(\frac{\lambda_1}{\lambda}\right)^{j_1} \cdots \left(\frac{\lambda_N}{\lambda}\right)^{j_N} \cdot \mathrm{e}^{j_1 T_{j_1} z} \cdots \mathrm{e}^{j_N T_{j_N} z}. \quad (30)$$

Finally expression (14) of $W_0(t)$, *per server*, becomes

$$W_0(t) = (1-\rho)\cdot\sum_{i=0}^{[t/T_N]}\ \sum_{\substack{j_1,\dots,j_N\\(j_1+\dots+j_N=i)}} \frac{i!}{j_1!\cdots j_N!}\cdot\left(\frac{\lambda_1}{\lambda}\right)^{j_1}\cdots\left(\frac{\lambda_N}{\lambda}\right)^{j_N}$$

$$\cdot\, e^{-A(t,j_1,\dots,j_N)}\cdot\frac{[A(t,j_1\cdots j_N)]^i}{i!},$$

$$\text{with } A(t,j_1,\dots,j_N) = \lambda\cdot(j_1T_1+\dots+j_NT_N-t) < 0, \quad (31)$$

$[t/T_N]$ is the integer part of (t/T_N). This expression is too long for numerical calculations. On the contrary, asymptotic expression (15) becomes

$$W_1(t) = 1 - \frac{1-\rho}{\rho_1\cdot e^{\beta_0T_1}+\dots+\rho_N\cdot e^{\beta_0T_N}-1}\cdot e^{-\beta_0t}, \quad (32)$$

where $q=\beta_0$ (>0) is the real (positive) root, closest to the origin, of the equation

$$q+\lambda-\lambda_1e^{qT_1}-\dots\cdot-\lambda_Ne^{qT_N}=0. \quad (33)$$

Finally, for $\rho<0.8$ and $P(\geq j)\leq 10^{-3}$, it is sufficient to use approximation (28) with expression (32), j being here the number of waiting customers just before an arrival:

$$P(\geq j)\cong I_2(j). \quad (34)$$

This quantity is useful to dimension the *buffer* controlling the queue. Each packet *occupies the buffer during the sojourn time* (= queueing delay + sending time). j packets waiting correspond to $(j+s)$ packets in the buffer. A packet is rejected on its arrival if the number j of packets waiting is such as $(j+s)\geq K$, K being the *buffer capacity*. Due to (34), the *rejection rate* R is

$$R\cong I_2(K-s). \quad (35)$$

Due to the resending of packets rejected, the traffic handled is not decreasing, and *for low values of* R, expression (35) is a *good approximation.*

4.3. An Example

As an example for the numerical calculations, we consider the case $N=3$; $T_1=1$, $T_2=5$, $T_3=30$; $\rho_1=\rho_2=\rho_3=0.2$ $(\rho=0.6)$. Following

TABLE I *Buffer capacity K* necessary for a *rejection rate R* in the case of *s servers*: formula (35) – (M/G/s queue)

s	R			
	10^{-3}	10^{-4}	10^{-5}	10^{-6}
1	28	35	39	44
2	28	38	49	58
3	28	38	49	59
5	28	38	49	59

Example (3 packet traffic streams): $N=3$; $T_1=1$, $T_2=5$, $T_3=30$; $\rho_1=\rho_2=\rho_3=0.2$ ($\rho=0.6$).

formula (35), Table I gives the buffer capacity necessary for $R=10^{-3}$, 10^{-4}, 10^{-5} and 10^{-6}, the multiserver capacity being $s=1$, 2, 3 and 5. As we can see, the buffer capacity K has to increase much when we want to decrease the rejection rate R. But an interesting result appears *for* $s>1$, *the buffer capacity does not depend on the multiserver capacity, approximately*. And from this Table I, it appears another interesting result: *for* $s\geq 1$, *to get a low rejection rate it is necessary to satisfy the condition:*

$$K>\frac{T_N}{T_1}. \tag{36}$$

References

[1] P. Le Gall, The stationary G/G/s queue with non-identical servers, *JAMSA* **11**: 2, 163–178 (1998).

[2] F. Pollaczek, Problèmes stochastiques posés par le phénomène de formation d'une queue d'attente à un guichet et par des phénomènes apparentés, *Mémorial des Sciences Mathématiques*, Gauthier-Villars, Paris CXXXVI (1957) (= GI/G/1 queue; in French).

[3] F. Pollaczek, Théorie analytique des problèmes stochastiques relatifs à un groupe de lignes télé-phoniques avec dispositif d'attente, *Mémorial des Sciences Mathématiques*, Gauthier-Villars, Paris CL (1961) (= GI/G/s queue; in French).

[4] N.U. Prabhu, Queues and Inventories, J. Wiley and Sons, New York (1965).