Journal of Applied Mathematics and Stochastic Analysis, **11**:2 (1998), 163-178.

THE STATIONARY G/G/s QUEUE WITH NON-IDENTICAL SERVERS

PIERRE LE GALL
France Telecom, CNET
4 Parc de la Bérengère
F-92210 Saint-Cloud, France

(Received February, 1998; Revised April, 1998)

We extend a recently developed *factorization method* to the case of the G/G/s queue with non-identical servers, by presenting *three simple properties* which lead to a simple numerical calculation method. We compare our results with those determined by classical Markovian (phase) methods in the case of the symmetrical M/G/s queue, and for the mean queueing delay we compare with results given by traffic simulation.

Key words: G/G/s Queue, GI/G/s Queue, First Come-First Served, Factorization, Singular Points.

AMS subject classifications: 60K25, 90B22.

1. Introduction

In a recent paper [3], we studied the stationary G/G/s queue by means of a new *factorization method* more general than a Wiener-Hopf type of decomposition. In Section 3, we show how this method may be readily extended to the case of non-identical servers for ***delayed*** *customers.* In Section 4, we then show how the effect of the busy period can be combined with that of the partial occupancies to evaluate the *probability of delay.* The calculations are made in both Sections 3 and 4 with the aim of deriving a *simple numerical calculation method* offered by ***three simple properties.*** We apply the results successively to the stationary GI/G/s queue and the M/G/s queue. In Section 5, we close with numerical comparisons of results obtained by applying classical Markovian methods to the symmetrical $M/E_2/s$ and $M/H_2/s$ queues, and we compare the calculated values of the mean queueing delay with results given by traffic simulation.

We begin, in Section 2, by defining notation and assumptions and by outlining the recently-derived preliminary results for the case of the symmetrical G/G/s queue.

2. Notation, Assumptions and Preliminary Results for the Symmetric G/G/s Queue

2.1 Notation and Assumptions

Except for the service time distribution, different for each server, the notations and assumptions will be the same as in Le Gall [3]. We consider a queue handled by a multiserver of **s** *non-identical servers.*

a) The Arrival Process: We assume a metrically transitive, strictly stationary process of successive non-negative interarrival times. Let $\boldsymbol{N(t)}$ denote *the random number of arrivals in the interval* $\boldsymbol{(0,t]}$. We write $\boldsymbol{dN(t)=1}$ or **0** depending on whether or not there is an arrival in the elementary interval $(t, t+dt)$. We *exclude the possibility of simultaneous arrivals.* We can then write

$$\mathrm{E}\{dN(t_0)\cdot dN(t_0+t)\} = \mathrm{E}[dN(t_0)]\cdot \boldsymbol{\rho(t)}\cdot dt, \tag{1}$$

where $\boldsymbol{\rho(t)}$ is the *arrival rate* at time $t+t_0$ if an arbitrary arrival occurred at time $\boldsymbol{t_0}$. We let

$$\int_0^\infty e^{zt}\cdot \rho(t)\cdot dt = \alpha_1(z) = \sum_{x=1}^{\infty} \varphi_{0,x}(z), \quad \mathbf{Re}(z)<0, \tag{2}$$

where $\varphi_{0,x}(z)$ corresponds to the xth arrival following the epoch t_0. However, the stationary assumption and the Abelian theorem give that $\lim_{z\to 0} z\cdot\alpha_1(z) = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the mean arrival rate. In a more general way, we may write for $j=1,2,\ldots$

$$\begin{aligned}&\mathrm{E}\{dN(t_0)\cdot dN(t_0+t_1)\ldots dN(t_0+t_1+\ldots+t_{j-1}+t_j)\}\\ &\qquad = [dN(t_0)]\cdot \boldsymbol{f_j(t_1\ldots t_j)}\cdot dt_1\ldots dt_j,\end{aligned} \tag{3}$$

and for $\mathbf{Re}(z_j)<0$ and $j=1,2,\ldots$

$$\int_0^\infty e^{z_1 t_1}\cdot dt_1 \ldots \int_0^\infty e^{z_j t_j}\cdot dt_j \cdot \boldsymbol{f_j(t_1\ldots t_j)} = \boldsymbol{\alpha_j(z_1\ldots z_j)}. \tag{4}$$

In the case of a ***renewal process***, the successive *arrival intervals* $\boldsymbol{Y_n}$ are *mutually independent* and identically distributed, and we let $\varphi_0(z) = \mathrm{E}[e^{zY_n}]$, for $\mathrm{Re}(z)<0$.

Expression (2) becomes

$$\alpha_1(z) = \frac{\varphi_0(z)}{1-\varphi_0(z)}, \tag{5}$$

and expression (4) becomes

$$\alpha_j(z_j\ldots z_j) = \alpha_1(z_1)\ldots\alpha_1(z_j). \tag{6}$$

In fact, we assume $\boldsymbol{\varphi_0(z)}$ to be *holomorphic* at the origin. From Paul Levy's theorem, we deduce that $\boldsymbol{\varphi_0(z)}$ *exists for* $\mathrm{Re}\,(z)<\delta$ *where* δ *is a positive real number.*

b) The Service Times: The successive service times T_n are *mutually independent* and independent of the arrival process. For server j ($j = 1,\ldots,s$) the service times $T_n(j)$ are identically distributed with a distribution function $F_1(t;j)$; and we let $\varphi_1(z;j) = \mathrm{E}[e^{zT_n(j)}]$, for $\mathrm{Re}(z) < 0$. We exclude the possibility of batch service and, consequently,

$$F_1(0;j) = F_1(+0;j) = 0. \tag{7}$$

We assume $\varphi_1(z;j)$ to be *holomorphic* at the origin. From Paul Levy's theorem, we deduce that $\varphi_1(z;j)$ *exists for* $\mathrm{Re}(z) \leq \delta$, *where* δ *is a real positive number.*

c) The Service Discipline: The servers are supposed to be *non-identical* with different service time distributions. But they are *indistinguishable* for the service discipline which is "*first come-first served.*"

d) The Traffic Handled: Loynes [4] demonstrated the existence of the stationary regime. In Section 3, we shall see that the non-identical servers are equivalent (during the busy period in the stationary regime) to different single servers handling the same value η for the *traffic intensity per server* (at any period), with the necessary and sufficient condition:

$$\eta < 1. \tag{8}$$

e) Queueing Delay: Since the term "*waiting time*" means "*sojourn time*" in Little's formula, for clarity we prefer to use the term "*queueing delay*" τ for the queueing process only and for an arbitrary customer.

f) 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 G/G/s Queue

Now, we outline the recent results that were presented in [3] for the case of the *symmetrical* G/G/s queue in equilibrium. To avoid very complicated calculations, Le Gall defined the *singular points* of the function $\mathbf{E}[e^{-q\tau}]$, with $\mathrm{Re}(q) > 0$, relating to the *queueing delay* τ for an arbitrary ***delayed*** customer. Secondly, Le Gall established conditions under which this function is *holomorphic*, these conditions being satisfied by a more general *factorization method* than the Wiener-Hopf type of decomposition. The results will very easily enable one to tackle the difficult case of non-identical servers for the evaluation of the queueing delays of *delayed* customers.

2.2.1 The singular points

For a ***delayed*** customer, the queueing delay in server j is denoted $w(j)$. The queueing delay τ of this customer is

$$\tau = \mathrm{Min}^{+}[w(1),\ldots,w(s)] = \mathrm{Max}[0,\min(w(1),\ldots,w(s))].$$

From an expression given by Pollaczek [6], we may write for $\mathrm{Re}(q) \geq 0$

$$\begin{cases} e^{-q\tau} = 1 - \dfrac{1}{(2\pi i)^s} \cdot \int_{+0} \exp(z_1 \cdot w(1)) \cdot \dfrac{dz_1}{z_1} \ldots \int_{+0} \exp(z_s \cdot w(s)) \cdot \dfrac{dz_s}{z_s} \cdot \dfrac{q}{q + \sum_{\nu=1}^{s} z_\nu}, \\ \text{with } \mathrm{Re}(q + \sum_{\nu=1}^{s} z_\nu) > 0. \end{cases} \tag{9}$$

In Le Gall [3], for the symmetrical case, Theorem 1 gives the *singular points* of the function $\mathrm{E}[e^{-q\tau}]$. These are the singular points of the following function with $\mathrm{Re}(q) < 0$, not holomorphic for $\mathrm{Re}(q) > 0$:

$$\begin{cases} \boldsymbol{G_s(q)} = 1 - \dfrac{1}{(2\pi i)^s} \cdot \int_{+0} \dfrac{dz_1}{z_1} \ldots \int_{+0} \dfrac{dz_s}{z_s} \cdot \dfrac{q}{q + \sum_{\nu=1}^{s} z_\nu} \cdot \dfrac{1}{\boldsymbol{R_s(z_1 \ldots z_s, q)}}, \\ \text{with } \mathrm{Re}(q + \sum_{\nu=1}^{s} z_\nu) > 0, \end{cases} \tag{10}$$

and

$$\boldsymbol{R_s(z_1 \ldots z_s; q)} = 1 + \sum_{\lambda=0}^{s-1} \binom{s}{\lambda} \cdot (-1)^{s-\lambda} \cdot \alpha_{s-\lambda}(q \ldots q) \cdot \prod_{j=\lambda+1}^{s} [\varphi_1(z_j) - 1], \tag{11}$$

where $\alpha_{s-\lambda}(q \ldots q)$ is defined by expression (4). The physical meaning may be perceived in the GI/G/s case, where, due to expression (6), we may write

$$\boldsymbol{R_s(z_1 \ldots z_s; q)} = 1 + \sum_{\lambda=0}^{s-1} \binom{s}{\lambda} \cdot (-1)^{s-\lambda} \cdot [\alpha_1(q)]^{s-\lambda} \cdot \prod_{j=\lambda+1}^{s} [\varphi_1(z_j) - 1],$$

or more simply

$$\boldsymbol{R_s(z_1 \ldots z_s; q)} = \prod_{j=1}^{s} [1 - \alpha_1(q) \cdot (\varphi_1(z_j) - 1)]. \tag{12}$$

The case j corresponds to the *server* j, assumed to be in isolation with an arrival process corresponding to $\alpha_1(q)$. For the G/G/s queue, instead of $\boldsymbol{R_s}$, we want to define a holomorphic function $\boldsymbol{V_s(z_1 \ldots z_s; q)}$ for $\mathrm{Re}(z_j) \geq 0$ $(j = 1,\ldots,s)$ and $\mathrm{Re}(q) \geq 0$, so that $\boldsymbol{V_s(0 \ldots 0; q)}$ has the same singular points as the function $\boldsymbol{G_s(q)}$ given by expressions (10) and (11).

2.2.2 The factorization method

Let

$$\boldsymbol{V_s(z_1 \ldots z_s; q)} = 1 - U_s(z_1 \ldots z_s; q). \tag{13}$$

In Le Gall [3] for the symmetrical case, we proved that the function U_s has to satisfy the following conditions of factorization:

We set

$$U_s(z_1 \ldots z_s; q) = \frac{(-1)^s \cdot \alpha_s(q \ldots q) \cdot \prod_{j=1}^{s} [\varphi_1(z_j) - 1]}{R_s(z_1 \ldots z_s, q)} \cdot \prod_{i=1}^{s} M_i(z_1 \ldots z_s; q), \tag{14}$$

where R_s is defined by expression (11) and U_s is holomorphic for $\mathrm{Re}(z_i) \geq 0$ $(i = 1, \ldots, s)$ and $\mathrm{Re}(q) \geq 0$, with the following ***conditions for*** M_i:

a) M_i is holomorphic for $\mathrm{Re}(z_i) < 0$, $i = 1, \ldots, s$;

b) $M_i(z_1 \ldots z_{i-1}, -q - \sum_{\nu=1}^{i-1} z_\nu, z_{i+1} \ldots z_s; q) \equiv 1.$

Then, we have

$$V_s(0 \ldots 0; q) = G_s(q),$$

where $G_s(q)$ is given by expressions (10) and (11).

Note the following facts.

1) Factorization (14) is still valid when we substitute

$$\varphi_1(z_j; j) \text{ for } \varphi_1(z_j) \tag{15}$$

in expressions (11) and (14), where $\varphi_1(z_j; j)$ has been defined in subsection (2.1.b).

2) To establish this factorization method, we had to use expression (22) in Le Gall [3] for $R(q) \geq 0$:

$$U_s(0 \ldots 0; q) = \frac{(-1)^s}{(2\pi i)^s} \cdot \int_{-0} \frac{dz_1}{z_1} \ldots \int_{-0} \frac{dz_s}{z_s} \cdot \frac{q}{q + \sum_{\nu=1}^{s} z_\nu} \cdot U_s(z_1 \ldots z_s; q). \tag{16}$$

2.2.3 The function U_s

For the symmetrical G/G/s queue, the function U_s has been defined in Le Gall [3]:

$$U_s(z_1 \ldots z_s; q) = \mathrm{Exp}\left\{ \frac{-1}{(2\pi i)^s} \cdot \int_{-0} \left[\frac{1}{q + \zeta_1} + \frac{1}{z_1 - \zeta_1} \right] d\zeta_1 \right.$$

$$\left. \ldots \int_{-0} \left[\frac{1}{q + \sum_{\nu=1}^{s-1} z_\nu + \zeta_s} + \frac{1}{z_s - \zeta_s} \right] d\zeta_s \cdot \log N_s\{\zeta_1 \ldots \zeta_s\} \right\}$$

with

$$R\left(q + \sum_{\nu=1}^{i-1} z_\nu + \zeta_i \right) > 0, \quad i = 1, \ldots, s. \tag{17}$$

The function N_s is given by

$$\left\{\begin{array}{c} \dfrac{1}{\boldsymbol{N_s(z_1 \ldots z_s)}} = \dfrac{A_s(z_1 \ldots z_s)}{B_s(z_1 \ldots z_s)}, \\ \text{with } A_s(z_1 \ldots z_s) = (-1)^s \left\{ \prod_{j=1}^{s} [\varphi_j(z_j) - 1] \right\} \\ \cdot \alpha_s \left(-z_1, -z_1 - z_2, \ldots, -\sum_{\nu=1}^{s} z_\nu \right), \text{ and} \\ B_s(z_1 \ldots z_s) = 1 + \sum_{\lambda=0}^{s-1} \binom{s}{\lambda} \cdot (-1)^{s-\lambda} \cdot \left\{ \prod_{j=\lambda+1}^{s} [\varphi_1(z_j) - 1] \right\} \\ \cdot \alpha_{s-\lambda} \left(-\sum_{\nu=1}^{\lambda+1} z_\nu, \ldots, -\sum_{\nu=1}^{s} z_\nu \right), \end{array}\right. \tag{18}$$

where $\alpha_j(z_1 \ldots z_j)$ is defined by expression (4).

3. The Delayed Customer

In this section, we consider the busy period only and a server's behavior during this period (= *congestion state*). There is generated (for server j) a queueing delay $w(j)$ for $j = 1, \ldots, s$, while the *multiserver* generates a *queueing delay* τ for an *arbitrary* ***delayed*** *customer*, in a *stationary regime.*

3.1 The Traffic Per Server

Let $T(j)$ denote the service time of an arbitrary customer *in server* j ($j = 1, \ldots, s$), in a stationary regime. The *termination rate* is $\mu_j = (1/T(j))$. The *total termination rate* for the *multiserver* (during a busy period in the steady state) is

$$\mu = \sum_{j=1}^{s} \mu_j = \left(\frac{s}{\bar{T}}\right), \tag{19}$$

defining the *mean service time* $\bar{T}$ *of the multiserver.* Let

$$\mu_j = \frac{\mu}{k_j} \quad \text{with} \sum_{j=1}^{s} \frac{1}{k_j} = 1. \tag{20}$$

In other words, k_j (> 0) *is defined by* the relation

$$\overline{T(j)} = k_j \cdot \left(\frac{\bar{T}}{s}\right). \tag{21}$$

The total arrival rate is $\lambda = [\mathrm{E}dN(t)]$. The traffic intensity handled by the multiserver is $\lambda \cdot \bar{T} = [\mathrm{E}dN(t)] \cdot \bar{T}$. Due to the stationary regime and expression (20) for μ_j, the arrival rate at *server* j is

$$\lambda_j = \frac{\lambda}{k_j} = \frac{1}{k_j} \cdot \mathrm{E}dN(t)]. \tag{22}$$

The *traffic intensity* handled by *server* j is, due to expressions (21) and (22),

$$\eta_j = \lambda_j \cdot \overline{T(j)} = \frac{1}{s} \cdot (\lambda \cdot \bar{T}) = \eta. \tag{23}$$

This traffic intensity is the same in each server. We may conclude that the following property holds.

Property 1 (Behavior of *server j* for the non-symmetrical G/G/s queue in a stationary regime): *Server* $\boldsymbol{j}$ behaves as a G/G/1 server, as if an arbitrary arrival is chosen with probability $(1/k_j)$ of being handled by *server* $\boldsymbol{j}$. It follows that the traffic intensity η has the same value in each server during a busy period, in the steady state.

Note the following facts.

1) The *symmetrical G/D/s queue* has to be excluded because of a deterministic mechanism for the choice of arrivals. But, when the service times are non-identical, the above property is correct.
2) For the *symmetrical G/G/s queue*, we have

$$\boldsymbol{k_j} = s, \quad j = 1,\ldots,s. \tag{24}$$

We deduce the *queueing delay* τ of an arbitrary ***delayed*** customer in the G/G/s queue by applying the expectation operator to expression (9) to evaluate $\mathrm{E}[e^{-q\tau}]$. But, we want to check in the case when $\boldsymbol{k_j}$ is valued in set $(1,2,\ldots,s)$.

3.2 The Distribution of the Queueing Delay

Expressions (9) and (10) may be applied immediately. For an arrival process, being renewal, expression (12) with substitution (15) becomes

$$R_s(z_1\ldots z_s;q) = \prod_{j=1}^{s} [1 - \alpha_1(q)\cdot(\varphi_1(z_j;j) - 1)]. \tag{25}$$

For a general stationary arrival process, as described in paragraph (2.1.a), we can substitute

$$\alpha_{s-\lambda}(q\ldots q) \text{ for } [\alpha_1(q)]^{s-\lambda} \tag{26}$$

after having expanded expression (25). Finally, with the new expressions for R_s and N_s, expressions (17) and (18) are still valid if we apply substitution (15). However, it may be very useful to note that the result is not changed if we replace

$$\alpha_{s-\lambda}\left(-\sum_{\nu=1}^{\lambda+1} z_\nu,\ldots,-\sum_{\nu=1}^{s} z_\nu\right)$$

in expression (18) by

$$\alpha_{s-\lambda}\left(-\sum_{\nu=1}^{k_{\lambda+1}} z_\nu,\ldots,\sum_{\nu=1}^{k_s} z_\nu\right).$$

To check this equivalence, we come back to the preceding expressions, related to (17) and (18) and leading to relations (30) through (33) in Le Gall [3], to be applied in the above expression (16) in order to satisfy the factorization method. It follows that the successive residues for $\zeta_i = z_i$ $(i = s, s-1,\ldots,1)$ lead to the application of $\alpha_{s-\lambda}$ to expression (16). The successive residues at the respective poles $z_i = -q - \sum_{\nu=1}^{i-1}$, for $i = s, s-1,\ldots,1$, lead to the expression $\alpha_{s-\lambda}(q\ldots q)$, as with the above expression for $\alpha_{s-\lambda}$. We find again expression (25) or a more general expres-

sion corresponding to expression (11). In other words, the singular points are not changed, and the factorization method is still applicable. The new expression for the functional $Ee^{-q\tau}$ does not change its value, since we know that the solution is unique to determine the G/G/s queue.

Now, we shall use the symmetry in expression (17), which leads to the following substitutions:

$$[\varphi_1(z_j;j)-1]\cdot\alpha_{s-\lambda}\left(-\sum_{\nu=1}^{k_{\lambda+1}}z_\nu,\ldots,-\sum_{\nu=1}^{k_j}z_\nu,\ldots,-\sum_{\nu=1}^{k_s}z_\nu\right)$$

$$=\frac{\varphi_1(z_j;j)-1}{z_j}\cdot(z_j)\cdot\alpha_{s-\lambda}\left(-\sum_{\nu=1}^{k_{\lambda+1}}z_\nu,\ldots,-\sum_{\nu=1}^{k_j}z_\nu,\ldots,-\sum_{\nu=1}^{k_s}z_\nu\right)$$

$$=\frac{\varphi_1(z_j;j)-1}{z_j}\cdot\left(\frac{1}{k_j}\cdot\sum_{\nu=1}^{k_j}z_\nu\right)\cdot\alpha_{s-\lambda}\left(-\sum_{\nu=1}^{k_{\lambda+1}}z_\nu,\ldots,-\sum_{\nu=1}^{k_j}z_\nu,\ldots,-\sum_{\nu=1}^{k_s}z_\nu\right).$$

Expression (16) leads to the expression:

$$\frac{\varphi_1(z_j;j)-1}{z_j}\cdot\left(\frac{q}{k_j}\right)\cdot\alpha_{s-\lambda}(q\ldots q).$$

Finally, considering every variable z_j $(j>\lambda)$, we find that (16) reduces to the expression:

$$\left[\prod_{j=\lambda+1}^{s}\left(\frac{q}{k_j}\right)\right]\cdot\alpha_{s-\lambda}(q\ldots q), \tag{27}$$

instead of $\alpha_{s-\lambda}(q\ldots q)$. Property 1 is satisfied and, finally, we may use expression (9).

Property 2 (The distribution of τ, the queueing delay of the non-symmetrical G/G/s queue): In a stationary regime, if $\boldsymbol{w(j)}$ is the queueing delay of an arbitrary ***delayed*** customer served by ***server j*** behaving under **Property 1**, the queueing delay of an arbitrary ***delayed*** customer served by a non-symmetrical (or symmetrical) ***multi-server*** is

$$\tau=\operatorname{Min}[w(1),\ldots,w(s)], \tag{28}$$

and the expression for the functional $\mathrm{E}[e^{-q\tau}]$ may be deduced from expression (9).

From the notation in formula (11) in [3], using Pollaczek's formula in [5], for $\operatorname{Re}(q)\geq 0$, we may write

$$\begin{cases} \mathrm{E}[e^{-qw(j)}]=\operatorname{Exp}\left\{\dfrac{-1}{2\pi i}\cdot\displaystyle\int_{+0}\left[\frac{1}{q+\zeta}-\frac{1}{\zeta}\right]\cdot\log\boldsymbol{K}(\zeta)\cdot d\zeta\right\}, \\ \text{with } \boldsymbol{K}(\zeta)=1-\dfrac{\alpha_1(-\zeta)}{k_j}\cdot[\varphi_1(\zeta;j)-1], \end{cases} \tag{29}$$

where $\boldsymbol{k_j}$ is defined by expression (21). In particular, when $|\boldsymbol{q}|$ increases indefinitely, we obtain, for the *probability of no delay* of this server equivalent to *server* $\boldsymbol{j}$ the

expression

$$Q_j = \mathrm{Exp}\left\{\frac{1}{2\pi i}\int_{+0} \log K(\zeta)\cdot\frac{d\zeta}{\zeta}\right\}, \tag{30}$$

and for the ***mean busy period size*** (= mean *number of customers* served) we deduce the expression

$$n_j = \frac{1}{Q_j}, \tag{31}$$

due to some classical relations. In fact, Q_j *is the probability of initiating a busy period.*

a) Case of the Non-Symmetrical GI/G/s Queue: Taking expressions (9), (28) and (29) into account, we have that the *queueing delay* τ, of the non-symmetrical GI/G/s queue for an *arbitrary* ***delayed*** *customer* is given [for $\mathrm{Re}(q) > 0$] by

$$\mathrm{E}[e^{-q\tau}] = 1 - \frac{1}{(2\pi i)^s}\cdot\int_{+0}\frac{dz_1}{z_1}\cdots\int_{+0}\frac{dz_s}{z_s}\cdot\left(\prod_{j=1}^{s}\mathrm{E}e^{-qw(j)}\right)\cdot\frac{q}{q+\sum_{\nu=1}^{s} z_\nu}. \tag{32}$$

Expression (35) below will give a simpler expression for the complementary distribution function of the queueing delay, which will be very convenient for numerical calculations.

b) Case of the Non-Symmetrical G/G/s Queue: For the non-symmetrical G/G/s queue, it is much more difficult to expand the terms between brackets in expression (32), since servers are mutually dependent through the arrival process. Finally, it is simpler to proceed in expressions (17) and (18) with the substitutions (15). But the expression in the distribution function of the queueing delay is much more intricate than that of the GI/G/s queue.

3.3 The Busy Period

The busy period corresponds to the s servers being *busy* simultaneously. To evaluate the probability Q_0 that the ***multiserver*** initiates a busy period, we may note that, among a great number N of successive arrivals, the mean number $N\cdot Q_0$ initiates a busy period of the multiserver and the mean number $N\cdot[Q_j/k_j]$ corresponds to server j due to **Property 1**. We deduce the relations:

$$Q_0 = \sum_{j=1}^{s}\frac{Q_j}{k_j} \quad \text{and} \quad \frac{1}{n} = \sum_{j=1}^{s}\frac{1}{n_j\cdot k_j}, \tag{33}$$

due to expression (31), with n being the *mean busy period size* of the multiserver, and k_j being defined by expression (21). In the case of a *symmetrical* ***G/G/s*** *queue*, where $k_j = s$ (see expressions (24)), we have

$$n_j = n, \quad j = 1,\ldots,s. \tag{34}$$

3.4 The Distribution Function

Let W the queueing delay of an ***arbitrary*** (delayed or not delayed) customer and let $F(t)$ denote the queueing delay distribution function of a ***delayed*** customer. Introduce the complementary function $G(t) = [1 - F(t)]$ for the multiserver and introduce

$G_j(t)$ for server j.

3.4.1 Case of the non-symmetrical G/G/s queue

Let $\mathcal{F}(t)$ denote the queueing delay distribution function of an *arbitrary* (delayed or not delayed) customer of the multiserver. We have

$$\mathcal{F}(t) = 1 - P \cdot G(t), \tag{35}$$

where P is the *probability of delay.*

3.4.2 Case of the non-symmetrical GI/G/s queue

As we consider the non-symmetrical GI/G/s queue, expression (28) makes say that

$$G(t) = \prod_{j=1}^{s} G_j(t), \tag{36}$$

since the relation $\tau > t$ needs simultaneously to have $w(j) > t$ for $j = 1, \ldots, s$. Expression (29) allows us to evaluate $G_j(t)$. Expressions (35) and (36) give for the moments of W:

$$\mathrm{E}(W^\alpha) = \alpha P \cdot \int_0^\infty t^{\alpha-1} \cdot G(t) \cdot dt, \quad \alpha = 1, 2, \ldots \tag{37}$$

$G_j(t)$ and this expression (37) can be easily calculated on a computer.

3.4.3 The non-symmetrical GI/M/s queue

From Pollaczek [5] we deduce from expression (29) that

$$G_j(t) = \mathrm{Exp}\left\{ -\frac{y_0(j)}{T(j)} \cdot t \right\}, \tag{38}$$

where $y_0(j)$ is the unique root, for $R(\zeta) > 0$, of

$$K(\zeta) = 1 - \frac{\alpha_1(-\zeta)}{k_j} \cdot [\varphi_1(\zeta; j) - 1] = 0, \quad \text{with } \varphi_1(\zeta; j) = \frac{1}{1 - \frac{\zeta}{T(j)}},$$

or

$$\frac{y_0(j)}{T(j)} = \frac{1}{1 + \frac{1}{k_j} \cdot \alpha[-y_0(j)]}. \tag{39}$$

Expressions (30) and (31) become

$$Q_j = y_0(j), \quad \text{and} \quad n_j = \frac{1}{y_0(j)}. \tag{40}$$

We deduce from expression (21) that

$$\frac{y_0(j)}{T(j)} = \frac{1}{n_j \cdot k_j} \cdot \frac{s}{\bar{T}}, \quad \text{with } \frac{s}{\bar{T}} = \sum_{j=1}^{s} \frac{1}{T(j)}. \tag{41}$$

Finally, expression (33) gives us that

$$\sum_{j=1}^{s} \frac{y_0(j)}{T(j)} = \frac{s}{n \cdot \overline{T}}, \tag{42}$$

where $\boldsymbol{n}$ is the *mean busy period size* of the multiserver (= mean number of customers served during a busy period). Consequently, we may write the following simple expression from expressions (36) and (38) for the complementary distribution function of the ***delayed*** customer:

$$G(t) = \mathrm{Exp}\left\{ -\frac{s}{n \cdot \overline{T}} \cdot t \right\}. \tag{43}$$

3.4.4 The non-symmetrical M/G/s queue

Due to expression (23), we may write

$$Q_j = 1 - \eta \quad \text{and} \quad n = n_j = \frac{1}{1-\eta}, \tag{44}$$

for the *mean busy period size* $\boldsymbol{n}$ of the multiserver. From Pollaczek [5], we may write

$$G_j(t) = \frac{1-\eta}{\eta} \cdot \sum_{\nu=1}^{\infty} \eta^{\nu} \cdot \left\{ 1 - \left[\int_0^t \frac{1 - F_1(u;j)}{\overline{T(j)}} \cdot du \right]^{(\nu)} \right\}, \tag{45}$$

where $[\cdot]^{(\nu)}$ denotes the k-fold convolution of the function $[\cdot]$. Finally, (36) gives an intricate expression for the complementary distribution function $\boldsymbol{G(t)}$ of the ***delayed*** customer:

$$G(t) = \prod_{j=1}^{s} \left\{ \frac{1-\eta}{\eta} \cdot \sum_{\nu=1}^{\infty} \eta^{\nu} \cdot \left(1 - \left[\int_0^t \frac{1 - F_1(u;j)}{\overline{T(j)}} \cdot du \right]^{(\nu)} \right) \right\}, \tag{46}$$

which can be evaluated numerically on a computer. To use expression (35) we, now, need to define the *probability of delay* $\boldsymbol{P}$.

4. The Probability of Delay

During the busy period (= the *congestion state*), server's behavior has been defined in a way quite independent of partial occupancy states. For these states, it follows that a *busy period appears exactly as a unique congestion state in the lost call model, with* $\boldsymbol{n}$ *successive service times handled as if there were a unique arrival*, with $\boldsymbol{n}$ being the mean value of the busy period size, i.e., of the number of customers served during this busy period. This fact could not be observed with the classical Markovian methods, and it has not been noted in Pollaczek's equation of [6].

With the lost call model in a stationary mode, let $P_i = P_0 \cdot h(i)$ denote the probability that i servers are busy upon the arrival of an arbitrary customer. The *probability of loss* is $P_a = P_s$ with

$$\frac{1}{P_a} = \frac{1}{P_s} = 1 + \frac{1 + \ldots + h(s-1)}{h(s)}, \tag{47}$$

due to the normalizing condition $\sum_{i=0}^{s} P_i = 1$.

With our preceding remark to evaluate the *probability of delay* P, we have to substitute

$$n \cdot h(s) \text{ for } h(s). \tag{48}$$

We may now conclude.

Property 3 (The probability of delay): For a non-symmetrical **G/G/s** queue in a stationary regime, the ***probability of delay*** is

$$P = \frac{n \cdot P_a}{1 + (n-1) \cdot P_a}, \tag{49}$$

where P_a is the ***probability of loss*** in the lost call model and n is the ***mean value of the busy period size*** as defined by expression (33).

As already seen in [1] and [2], we know that the evaluation of P_a is extremely difficult except in two symmetrical cases: The GI/M/s and M/G/s queues. In particular, for the **M/G/s** queue we conclude that *the delay Erlang formula may be extended for a general service time distribution.* In that case, expression (44) gives $n = [1/(1-\eta)]$. From fact 1) after **Property 1**, we know that the **M/D/s** queue has to be excluded; however, it has already been noted by C. Palm that the delay Erlang formula gives still an excellent approximation.

5. Numerical Comparisons for Some M/G/s Queues

Now, we present some numerical comparisons with the results obtained by applying classical Markovian methods to the symmetrical $M/E_2/s$ and $M/H_2/s$ queues.

5.1 The Symmetrical $M/E_2/s$ Queue

The service time distribution of any server j is

$$\begin{cases} F_1(t) = 1 - e^{-2t} \cdot (1+2t); & \varphi_1(z) = \dfrac{1}{(1-\frac{z}{2})^2}; \\ \text{corresponding to } \mathbf{E(T) = 1}; & C_s^2 = \dfrac{\mathrm{Var}T}{[\mathrm{E}(T)]^2} = \mathbf{0.5}. \end{cases} \tag{50}$$

We deduce, for the queueing delay $w(j)$ of any server j when $\mathrm{Re}(z) < 0$:

$$\begin{cases} \mathrm{E}[e^{zw(j)}] - \dfrac{1-\eta}{1-\eta \cdot \frac{\varphi_1(\xi)-1}{\xi}} = \dfrac{(1-\eta)(1-\frac{z}{2})^2}{(1-\frac{z}{2})^2 - \eta \cdot (1-\frac{z}{4})} = \dfrac{4(1-\eta)(1-\frac{z}{2})^2}{(z-\beta_1)(z-\beta_2)}, \\ \text{with } \beta_1 = 2(1-\frac{\eta}{4}) - 2\sqrt{\frac{\eta}{2} \cdot \left(\frac{\eta}{8}+1\right)}, \\ \text{and } \beta_2 = 2\left(1-\frac{\eta}{4}\right) + 2\sqrt{\frac{\eta}{2} \cdot \left(\frac{\eta}{8}+1\right)}. \end{cases} \tag{51}$$

The queueing delay distribution function of this server j is

$$\begin{cases} F_j(t) = F_1(t) = 1 - \eta \cdot G_1(t), \\ \text{with } G_1(t) = (1-\eta) \cdot \left\{ \dfrac{1}{(\beta_2 - \beta_1)} \cdot \dfrac{4-\beta_1}{\beta_1} \cdot e^{-\beta_1 t} + \dfrac{1}{(\beta_1 - \beta_2)} \cdot \dfrac{4-\beta_2}{\beta_2} \cdot e^{-\beta_2 t} \right\}. \end{cases} \tag{52}$$

From expression (36), for the multiserver, the queueing delay distribution function of the *delayed* customer is

$$p = F(t) = 1 - [G_1(t)]^s. \tag{53}$$

where t is the *conditional queueing delay percentile* $t(p)$.

In Table 1, we give (for $\eta = 0.8$) our results concerning $t(p)$ and the probability of delay P from expression (49), for $s = 2, 5, 10, 25, 50$ and $p = 0.5, 0.9, 0.95, 0.99$. For the results given by the Markovian methods (= *phase* methods), we refer to Table 1 (first part) in Seelen and Tijms [7]. These results appeared as approximated (Markovian) results in our Table 1. The deviation is not significant.

	p	0.5	0.9	0.95	0.99	P
	exact	1.36	4.31	5.58	8.52	0.711
$s = 2$	Markov	1.34	4.29	5.55	8.50	0.709
	exact	0.58	1.78	2.29	3.47	0.554
$s = 5$	Markov	0.55	1.73	2.24	3.42	0.548
	exact	0.31	0.93	1.19	1.78	0.409
$s = 10$	Markov	0.29	0.88	1.13	1.72	0.402
	exact	0.13	0.40	0.51	0.75	0.209
$s = 25$	Markov	0.12	0.36	0.47	0.70	0.203
	exact	0.07	0.21	0.27	0.40	0.087
$s = 50$	Markov	0.06	0.19	0.24	0.36	0.084

Table 1: The symmetrical $M/E_2/s$ queue

1) **Results** a) the conditional queueing delay percentile $t(p)$ for the ***delayed*** customer
   b) probability of delay: P
2) **Parameters** a) traffic intensity per server: $\eta = 0.8$
   b) service time distribution from expression (50): $E(T) = 1$, $C_s^2 = 0.5$
3) **Calculations** a) "*exact*": Section 5.1
   b) "*Markov*": Phase method

5.2 The Symmetrical $M/H_2/s$ Queue

The service time distribution of any server j is:

$$\begin{cases} F_1(t) = 0.5 \cdot [1 - e^{-b_1 t}] + 0.5 \cdot [1 - e^{-b_2 t}], \; b_1 = \dfrac{2}{2-\sqrt{2}}; \; b_2 = \dfrac{2}{2+\sqrt{2}}, \\ \text{corresponding to } E(T) = 1, \quad C_s^2 = \dfrac{\mathrm{Var} T}{[E(T)]^2} = 2. \end{cases} \tag{54}$$

We let

$$\begin{cases} \beta_1 = \dfrac{b_1 + b_2 - \eta}{2} + \sqrt{\dfrac{(b_1 + b_2 - \eta)^2}{4} - b_1 \cdot b_2 \cdot (1 - \eta)}, \\ \beta_2 = \dfrac{b_1 + b_2 - \eta}{2} - \sqrt{\dfrac{(b_1 + b_2 - \eta)^2}{4} - b_1 \cdot b_2 \cdot (1 - \eta)}. \end{cases} \tag{55}$$

As for expression (52), the complementary queueing delay distribution function of any server j is

$$G_1(t) = \frac{\beta_2 - 1 + \eta}{\beta_2 - \beta_1} \cdot e^{-\beta_1 t} + \frac{\beta_1 - 1 + \eta}{\beta_1 - \beta_2} \cdot e^{-\beta_2 t}. \tag{56}$$

Now, table 2 uses expressions (53) and (56) to give the new values of $t(p)$ and P corresponding to the same values of parameters η, s and p as in Table 1. For the approximated (Markovian) results we refer to Table 1 (second part) in Seelen and Tijms [7]. The deviation is not significant either.

	p	0.5	0.9	0.95	0.99	P
	exact	2.47	8.65	11.30	17.43	0.711
$s = 2$	Markov	2.36	8.82	11.62	18.11	0.715
	exact	0.89	3.35	4.41	6.86	0.554
$s = 5$	Markov	0.89	3.41	4.53	7.12	0.562
	exact	0.40	1.58	2.11	3.34	0.409
$s = 10$	Markov	0.43	1.63	2.19	3.48	0.418
	exact	0.15	0.56	0.76	1.22	0.209
$s = 25$	Markov	0.16	0.60	0.81	1.31	0.216
	exact	0.07	0.26	0.34	0.55	0.087
$s = 50$	Markov	0.08	0.28	0.38	0.61	0.090

Table 2: The symmetrical $M/H_2/s$ queue

1) **Results** a) the conditional queueing delay percentile $t(p)$ for a ***delayed*** customer
   b) probability of delay: P
2) **Parameters** a) traffic intensity per server: $\eta = 0.8$
   b) service time distribution from expression (54): $E(T) = 1$, $C_s^2 = 2$
3) **Calculations** a) "*exact*": Section 5.2
   b) "*Markov*": Phase method

Tables 3 (for $\eta = 0.5$) and 4 (for $\eta = 0.9$) give some comparisons between values of the *mean queueing delay* $\bar{W}$ of an *arbitrary* (delayed or not delayed) customer, given by calculation and by traffic simulation. Now, the service time distribution of any server j is

$$\begin{cases} F_1(t) = a_1 \cdot [1 - e^{-b_1 t}] + (1 - a_1) \cdot [1 - e^{-b_2 t}], \\ \text{with } a_1 = \dfrac{50 \cdot (n-1)}{50 \cdot n + 31}, \quad b_1 = 10, \quad b_2 = \dfrac{10 \cdot (1 - a_1)}{10 - a_1}, \\ \text{corresponding to } \mathrm{E}(T) = 1, C_s^2 = \dfrac{\mathrm{Var}\, T}{[\mathrm{E}(T)]^2} = n. \end{cases} \tag{57}$$

	n	5	10
$s = 2$	C	0.98	1.76
	S	1.0	1.77
$s = 5$	C	0.14	0.24
	S	0.15	0.25

Table 3: The symmetrical $M/H_2/s$ queue
The ***mean queueing delay*** $\bar{W}$ of an *arbitrary* customer

1) **Comparisons** between calculations (expression (58)), line C and simulations, line **S**.
2) **Parameters** a) traffic intensity per server: $\eta = 0.5$
b) service time distribution from expression (57): $\mathrm{E}(T) = 1, C_s^2 = n$

	n	5	10
$s = 5$	C	4.5	8.3
	S	4.6	8.7
$s = 10$	C	1.9	3.3
	S	2.0	3.2

Table 4: The symmetrical $M/H_2/s$ queue
The ***mean queueing delay*** $\bar{W}$ of an *arbitrary* customer

1) **Comparisons** between calculations (expression (58)), line C and simulations, line **S**.
2) **Parameters** a) traffic intensity per server: $\eta = 0.9$
b) service time distribution from expression (57): $\mathrm{E}(T) = 1, C_s^2 = n$

On applying expressions (55) and (56), we deduce from (37):

$$\bar{W} = P \cdot \int_0^\infty [G_1(t)]^s \cdot dt, \tag{58}$$

where P is given by expression (49). Tables 3 and 4 consider the cases $n = 5$ and 10. Taking the accuracy of simulations and of calculations of (58) into account, the results given by traffic simulations and by calculations are in good agreement.

6. Conclusion

We characterized non-symmetrical (and symmetrical) G/G/s queues by ***three simple properties*** which lead to very simple numerical calculations, at least for the GI/G/s queues. The deviation of the numerical results with those of classical, approximated, Markovian methods is not significant, and our numerical results are in good agreement with those given by traffic simulations.

References

[1] Le Gall, P., General telecommunications traffic without delay, *Proc. International Teletraffic Congress*, ITC **8** (Melbourne, Australia, Nov. 1976) 125.

[2] Le Gall, P., Sur le problème du trafic téléphonique général, direct et sans attente, *Annales Telec.*, Paris (Sept.-Oct.) **34**:9-10 (1979), 459-468.

[3] Le Gall, P., The stationary G/G/s queue, *JAMSA* **11**:1 (1998), 59-71.

[4] Loynes, R.M., The stability of a queue with nonindependent interarrival and service times, *Proc. Cambridge Philos. Soc.* **58** (1962), 494-520.

[5] Pollaczek, F., 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).

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

[7] Seelen, L.P. and Tijms, H.c., Approximations to the waiting time percentiles in the M/G/c queue, *Proc. Intern. Teletraffic Cong.* ITC-11 (Kyoto), Sept. 1985, 1.4-4.