Profiting the idleness in single server system with orbit-queue

Amar Aissani
University of Science and Technology Houari Boumediene (USTHB)
BP 32 El Alia, Bab-Ez-Zouar, Algiers 16 111, Algeria
aaissani@usthb.dz

Tuan Phung-Duc
Faculty of Engineering, Information and Systems
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
tuan@sk.tsukuba.ac.jp

ABSTRACT

How Profitable is the period of idleness of a server in a network? Queueing theory aims to answer to such a question amongst others. We restrict ourself to the problem of profitability of idleness of the server from the point of view of two-way communication in call centers in relation with retrial queueing models. This problem has been investigated with the pure Markovian setting in both cases of linear and constant retrial policies. The present work aims to extend the Markovian model to the non-Markovian one's for constant retrial policy in the case of general service times probability distributions.

CCS Concepts

•Networks → Network performance modeling;

Keywords

Profiting idleness; M/G/1; Two-way communication, Orbit-queue

1. INTRODUCTION

How Profitable is the period of idleness of a server in a network? Queueing theory aims to answer to such a question amongst others (allocation of resources and storage capacities before service). In classical models, primary arrival customer (e.g. request or call) finding the server blocked (busy or out of order) joins a queue with a given service discipline (FIFO, LIFO or RANDOM), or is considered to be a lost unit (Erlang model). When the server is busy, the arrival is not definitively lost, but is allowed to join a retrial group or an "orbit" (a sort of queue for secondary sources) and repeat successively his attempt until the server is able to provide service. Otherwise, if the server is available, the arriving customer begins service immediately. Such models are called systems with repeated calls or retrial queueing systems (see [4, 10]). These queueing models have been used in modelling Switching Networks, Digital Cellular Mobile Networks, Local Area Networks under the protocol of random multiple access and so on.

We consider here the two-way communication retrial queue with balanced call blending introduced in [5, 11] in relation with performance evaluation of call centers. Call blending consists in the mixing of incoming and outgoing call activities. Indeed, an important question in such a call center is whether it handles inbound traffic, with incoming calls, or outbound traffic, with outgoing calls. Usually, incoming calls are assigned to operators by an Automatic Call Distributor (CAD), while outgoing calls are either initiated by the CAD (automatically), or by the operators (manually). This mixture of incoming and outgoing calls is named call blending, with two-way communication, serving several purposes. It seems that retrial queueing models are adequate model according to the following principle [5]: (i) In the case of inbound call centers, operators (servers) can add to their regular tasks, outgoing calls in order to perform secondary non-urgent activities; (ii) It may occurs that both incoming and outgoing calls are vital elements of the service delivered in call center, and both should be delivered.

Several way of call blending can be identified and the reader is referred to the discussions in [9] which gives an excellent overview and other references (see [11]) about a high-level discussion and basic performance analysis and for a data-base of Markov chain models of call centers, including a discussion of model fidelity and efficiency in a simulation context. However, we concentrate on the idea of [5, 11] that such a principle of mixing can be analytically modeled by retrial queues supporting two way communication. In those works, the authors assume a "linear retrial policy" for incoming calls (i.e. the probability of retrial request during an infinitesimally small interval $(0, \Delta t)$ is proportional to the number m of customers in the queue (orbit) i.e. $m\mu\Delta t + 0(\Delta t)$. Such a choice is motivated by the fact that outgoing call activity is largest when few incoming calls are in orbit (the queue), and smallest when they are many in orbit.

The work [11] reconsiders this problem with the "constant retrial policy" in which the probability of a retrial request during an infinitesimally small interval $(0, \Delta t)$ is independent of the number of customers in the queue (orbit) i.e. $\mu\Delta t + 0(\Delta t)$. The choice of a constant retrial policy rather than a linear one gave rise to fierce controversy. However, if the linear one is "naturally" adapted to telephonic system and similar ones, the constant one seems to be adapted

VALUETOOLS 2017, December 5–7, 2017, Venice, Italy

ISBN 978-1-4503-6346-4/17/12…$15.00

DOI: https://doi.org/10.1145/3150928.3150929

to some modern communication systems in which we can "control" the retrial rate which is not possible in telephonic systems in which the linear "policy" is more adapted. For more discussion about the "balance" on the choice between "linear", "constant" or "general" policies, see the discussion in [5, 11] etc. The model with constant retrial rate perfectly fits for service systems including call centers with call-back option. Customers who are not served immediately upon arrival has an option to register their phone number (in call centers) or ticket number (in other service systems). When the server is idle it will call to registered customers or those with a ticket number. In these applications, the "retrial" is better interpreted as "call for customers". Since such kind of service systems is now ubiquitous in various social service systems such as hospital, ticket office in stations etc., analysis of such a system is very important from a practical point of view. In any case, the different "policies" conduct to different mathematical developments, results and interpretations. So, managers should be aware of when operating on these results.

We must focalize on the fact that there is some similarities and differences between "retrial" and "vacation" policies from the "mathematical" point of view and from the "management" point of view.

In the last one, the non-urgent tasks are provided when the server and all the system are idle (free of customers), i.e. at the end of a busy period and the beginning of an idle one. In our case, "profitability period" is taken each time the server is idle (and perhaps the system not). Another difference is that such an idle period of the server (and not of the system) occurs at random, while in traditional vacation models this "profitability" period is programmed when both the server and the system are idle.

Some authors [1, 3] have conducted some studies to emphasize mathematically these two physical phenomena: "retrial" (linear and/or constant) and "vacation" simultaneously. In any case, these studies show the mathematical similarities and differences between the two modelling concepts. They show also how "managers" can exploit the mathematical results for performance evaluation and optimization purposes.

Now, we restrict ourself to the problem of profitability of idleness of the server from the point of view of two-way communication in call center. This problem has been introduced in [5, 6] with the linear retrial policy in the particular case of exponential probability distributions (PDF) of service times of both incoming and outgoing calls. An extension to the case of general service's probability distributions is provided in [2, 13, 14, 7] with a discussion about the optimization of such an idle time when it has a sense. The present work aims to extend the model of [5, 11] for constant retrial policy again in the case of general service time probability distributions.

The organization of the paper is as follows. In the next section, we present the model and mathematical assumptions of the two-way communication retrial queue with balanced call blending in the context of "constant" retrial policy as discussed in [5, 11]. Section 3 is devoted to the derivation of the system of steady-state equations. In section 4, we obtain the partial generating functions of the joint distribution of the server state and orbit size. In Section 5, we derive the unconditional generating function of the number of customers in orbit (the queue in our case). As a consequence, we obtain the probabilistic performance measures (Section 6) and mean performance measures (Section 7). Numerical illustrations are provided in Section 8 showing how managers can exploit the obtained results.

2. MATHEMATICAL FORMULATION

We consider a single server retrial queue with two-way communication. Primary incoming call requests arrive at the server (or operator) according to a Poisson process with rate λ. Incoming calls finding an idle server receive service immediately. If such an incoming call finds the server busy, it leaves the server (but not the system as in the case of Erlang's loss model) for a random period of time (we say that it enters orbit, a sort of queue) "controlled" by a constant retrial policy with rate $\mu(1-\delta_{0,m})$ when there are m customers in the queue. Here, $\delta_{0,m}$ denotes the Kronecker symbol i.e.

$$\delta_{0,m} = \begin{cases} 0, & \text{if} \quad m = 0, \\ 1, & \text{if} \quad m \neq 0. \end{cases}$$

This means that the "denial requests" may retry their call at random time exponentially distributed with a rate independent of the number of customers in the system. With such a constant retrial policy, customers form a FIFO queue in the orbit, and only the customer at the head of the queue can request service. When the server turns idle, it stays idle for an exponentially distributed time with mean $1/\alpha$ (waiting for an incoming call from the orbit or a primary call) and then makes an outgoing call if there is not an incoming call arriving in the idle period.

The service of incoming and outgoing calls are i.i.d., arbitrary distributed with probability distributions functions $H_1(x)$ and $H_2(x)$ respectively. We denote respectively by $h_1(s)$ and $h_2(s)$, $Re(s) \geq 0$ the Laplace-Stieltjes transforms of these probability distributions, by h_1 and h_2 the first moments of service time and by h_{1n} and h_{2n} the respective n-th ($n \geq 2$) moments.

Let $S(t)$ be the server state at time t,

$$S(t) = \begin{cases} 0, & \text{if the server is idle,} \\ 1, & \text{if the server is providing an incoming service,} \\ 2, & \text{if the server is providing an outgoing service,} \end{cases}$$

and let $N(t)$ be the number of calls in orbit at time t. Additionally, we introduce a continuous random variable $\xi(t)$ which represents the residual service time if $S(t) = 1$ or $= 2$; $\xi(t) = 0$ if $S(t) = 0$.

From general theory [8], it is easy to see that the three-dimensional stochastic process

$$\zeta(t) = \{(S(t), N(t, \xi(t)); t \geq 0\}$$

defined on the state space $E = \{0, 1, 2\} \times \mathbb{N} \times \mathbb{R}^+$ is now a Markov process. It should be noted that $\mathbb{N}$ and $\mathbb{R}^+$ represent the set of non-negative integers and the set of non-negative real numbers, respectively.

3. SYSTEM OF STEADY-STATE EQUATIONS

The analysis of limiting behavior of the queueing process $\{\zeta(t); t \geq 0\}$ at a stationary point of time can be performed with the help of the Kolmogorov equations under some regularity conditions provided the limits

$$P_0(m) = \lim_{t\to\infty} P\{S(t) = 0, N(t) = m\}; \quad m = 0, 1, 2, ...,$$

$$P_i(m, x) = \lim_{t\to\infty} P\{S(t) = i, N(t) = m; \xi(t) < x\}, \quad i = 0, 1, 2; m = 0, 1, 2, ...; x \geq 0,$$

exist and are independent of the initial state. We can show that the steady-state probabilities exist if and only if the stability condition

$$\phi = \frac{\lambda+\mu}{\mu}\rho + \frac{\lambda}{\mu}\sigma < 1. \quad (1)$$

is fulfilled, where $\rho = \lambda h_1$ and $\sigma = \alpha h_2$. The necessary part of this condition appears in the following developments. So, the condition $\phi < 1$ is assumed to hold from now.

By usual arguments of Markov processes [8], we have the following system of Kolmogorov equations of statistical equilibrium,

$$(\lambda+\alpha+\mu(1-\delta_{0,m}))P_0(m) = \frac{dP_1(m,0)}{dx} + \frac{dP_2(m,0)}{dx}, m \geq 0, \quad (2)$$

$$\lambda P_1(m,x) = \frac{dP_1(m,x)}{dx} - \frac{dP_1(m,0)}{dx} + \lambda(1-\delta_{0m})P_1(m-1,x) + \lambda P_0(m)H_1(x) + \mu P_0(m+1)H_1(x), m \geq 0, \quad (3)$$

$$\lambda P_2(m,x) = \frac{dP_2(m,x)}{dx} - \frac{dP_2(m,0)}{dx} + \lambda(1-\delta_{0m})P_2(m-1,x) + \alpha P_0(m)H_2(x), m \geq 0 (4)$$

The above set of equations needs to be solved under the normalizing equation

$$\lim_{x\to\infty} \sum_{m=0}^{\infty}\{P_0(m) + \sum_{i=1}^{2} P_i(m,x)\} = 1.$$

4. SOLUTION OF THE SYSTEM OF EQUATIONS

In order to solve the system (2)-(4), we introduce the partial generating functions

$$Q_0(z) = \lim_{t\to\infty} E(z^{N(t)}; S(t) = 0) = \sum_{m=0}^{\infty} z^m P_0(m), i = 0, 1, 2; | z | \leq 1,$$

$$Q_i(z,x) = \lim_{t\to\infty} E(z^{N(t)}; S(t) = i, \xi(t) < x) = \sum_{m=0}^{\infty} z^m P_i(m,x), i = 1, 2; | z | \leq 1,$$

which are convergent for each $0 \leq x < \infty$ at least in the domain $| z | \leq 1$.

Multiplying equations (2)-(4) by z^m and summing on m from zero to ∞, we get

$$(\lambda+\alpha+\mu)Q_0(z) - \mu P_{00} = \frac{\partial Q_1(z,0)}{\partial x} + \frac{\partial Q_2(z,0)}{\partial x}, \quad (5)$$

$$(\lambda - \lambda z)Q_1(z,x) = \frac{\partial Q_1(z,x)}{dx} - \frac{\partial Q_1(z,0)}{\partial x} + \left(\lambda + \frac{\mu}{z}\right)Q_0(z)H_1(x) - \frac{\mu}{z}P_{00}H_1(x), | z | \leq 1, x \geq 0, \quad (6)$$

$$(\lambda - \lambda z)Q_2(z,x) = \frac{\partial Q_2(z,x)}{dx} - \frac{\partial Q_2(z,0)}{\partial x} - \alpha Q_0(z)H_2(x), \quad | z | \leq 1, x \geq 0. \quad (7)$$

Define now the Laplace transform

$$f_i(z,s) = \int_0^\infty e^{-sx}Q_i(z,x)dx, i = 1, 2,$$

which are convergent at least in the domain $| z | \leq 1$, $Re(s) > 0$. Applying these transforms to the equations (6)-(7), we obtain the following equations

$$s(s-\lambda+\lambda z)f_1(z,s) = \frac{\partial Q_1(z,0)}{\partial x} - \left[\left(\lambda+\frac{\mu}{z}\right)Q_0(z) - \frac{\mu}{z}P_{00}\right]h_1(s), \quad (8)$$

$$s(s-\lambda+\lambda z)f_2(z,s) = \frac{\partial Q_2(z,0)}{\partial x} - \alpha Q_0(z)h_2(s). \quad (9)$$

In order to find the function $\frac{\partial Q_1(z,0)}{\partial x}$ we use the fact that the function $f_2(z,s)$ is an analytic function in s in the domain $Re(s) > 0$. Since for $s = \lambda - \lambda z$ the right hand side of equation (8) vanishes, then the right hand side of this equation must be zero at this point. Thus we have

$$\frac{\partial Q_1(z,0)}{\partial x} = \left[\left(\lambda + \frac{\mu}{z}\right)Q_0(z) - \frac{\mu}{z}P_{00}\right]h_1(\lambda - \lambda z). \quad (10)$$

Similarly, for $s = \lambda - \lambda z$ the right hand side of (9) is equal to zero.

$$\frac{\partial Q_2(z,0)}{\partial x} = \alpha Q_0(z)h_2(\lambda - \lambda z). \quad (11)$$

By substitution relations (10) and (11) into (5), we obtain

$$Q_0(z) = \frac{\mu - \frac{\mu}{z}h_1(\lambda-\lambda z)}{\lambda+\mu-\left(\lambda+\frac{\mu}{z}\right)h_1(\lambda-\lambda z) + \alpha - \alpha h_2(\lambda-\lambda z)} P_{00}. \quad (12)$$

Now, the functions $f_1(z,s)$ and $f_2(z,s)$ are entirely defined from(8)-(12) up to the undetermined constant P_{00}:

$$f_1(z,s) = \frac{\left[(\lambda+\frac{\mu}{z})Q_0(z) - \frac{\mu}{z}P_{00}\right][h_1(\lambda-\lambda z) - h_1(s)]}{s(s-\lambda-\lambda z)}, \quad (13)$$

$$f_2(z,s) = \frac{\alpha Q_0(z)[h_2(\lambda-\lambda z) - h_2(s)]}{s(s-\lambda+\lambda z)}. \quad (14)$$

From (13)-(14) and using Tauberian theorems, we can derive the following expressions:

$$Q_1(z,\infty) = \lim_{s\to 0+} sf_1(z,s) = \frac{[(\lambda+\frac{\mu}{z})Q_0(z) - \frac{\mu}{z}P_{00}][1 - h_1(\lambda-\lambda z)]}{\lambda - \lambda z}, \quad (15)$$

$$Q_2(z,\infty) = \lim_{s\to 0+} sf_2(z,s) = \alpha\frac{1-h_2(\lambda-\lambda z)}{\lambda-\lambda z}Q_0(z). \quad (16)$$

By using normalization condition

$$Q_0(1) + Q_1(1,\infty) + Q_2(1,\infty) = 1,$$

we can get the undetermined constant P_{00}

$$P_{00} = \frac{1 - \frac{\lambda+\mu}{\mu}\rho - \frac{\lambda}{\mu}\sigma}{1+\sigma}. \quad (17)$$

where $\rho = \lambda h_1$ and $\sigma = \alpha h_2$. Because $P_{00} > 0$, it follows from (17) that a necessary condition for the stability condition of the system is given by

$$\phi = \frac{\lambda+\mu}{\mu}\rho + \frac{\lambda}{\mu}\sigma < 1.$$

It can be shown that it is visibly also sufficient.

5. PROBABILITY DISTRIBUTION OF THE ORBIT SIZE

It is not difficult to see that the probability distribution of the number of customers in orbit is

$$\lim_{t\to\infty} P\{N(t) = m\} = P_0(m) + P_1(m,\infty) + P_2(m,\infty).$$

So, the same relation yields with the generating functions due to their linear property

$$Q(z) = \lim_{t\to\infty} E\{z^{N(t)}\} = Q_0(z) + Q_1(z,\infty) + Q_2(z,\infty).$$

Multiplying the m general term by z^m and summing on m from zero to ∞ we obtain

$$\begin{aligned}
Q(z) &= Q_0(z) + Q_1(z,\infty) + Q_2(z,\infty) = \\
&= Q_0(z) + \left[\frac{(\lambda+\frac{\mu}{z})Q_0(z) - \frac{\mu}{z}P_{00}}{\lambda-\lambda z}\right][1-h_1(\lambda-\lambda z)] \\
&+ \alpha\frac{1-h_2(\lambda-\lambda z)}{\lambda-\lambda z}Q_0(z) = \\
&= \left[1 + (\lambda+\frac{\mu}{z})\frac{1-h_1(\lambda-\lambda z)}{\lambda-\lambda z} + \alpha\frac{1-h_2(\lambda-\lambda z)}{\lambda-\lambda z}\right]Q_0(z) \\
&- \frac{\mu}{z}\frac{1-h_1(\lambda-\lambda z)}{\lambda-\lambda z}P_{00}.
\end{aligned}$$

Finally, we obtain

$$Q(z) = \frac{1 - \frac{\lambda+\mu}{\mu}\rho - \frac{\lambda}{\mu}\sigma}{1+\sigma} \times \frac{\Phi(z)\big[\mu - \frac{\mu}{z}h_1(\lambda-\lambda z)\big] - \Psi(z)\frac{\mu}{z}[1-h_1(\lambda-\lambda z)]}{(\lambda-\lambda z)\Psi(z)} \quad (18)$$

where

$$\Phi(z) = \lambda - \lambda z + \big(\lambda+\frac{\mu}{z}\big)[1-h_1(\lambda-\lambda z)] + \alpha[1-h_2(\lambda-\lambda z)]$$

and

$$\Psi(z) = \lambda+\mu+\alpha - \big(\lambda+\frac{\mu}{z}\big)h_1(\lambda-\lambda z) - \alpha h_2(\lambda-\lambda z).$$

6. PROBABILISTIC PERFORMANCE METRICS

Now, using L'Hospital rule or Taylor expansion method [12] in (12), (15) and (16), we get the server state probabilities:

- the probability that the server is idle
$$M_0^0 = Q_0(1) = \frac{\mu(1-\rho)}{\mu-(\lambda+\mu)\rho-\lambda\sigma}P_{00} = \frac{1-\rho}{1+\sigma}. \quad (19)$$
- the probability that the server is busy by an incoming call
$$M_0^1 = Q_1(1,\infty) = h_1[(\lambda+\mu)Q_0(1,\infty) - \mu P_{00}] = \rho. \quad (20)$$
- the probability that the server is busy by an outgoing call
$$M_0^2 = Q_2(1,\infty) = \sigma Q_0(1,\infty) = \sigma\frac{1-\rho}{1+\sigma}. \quad (21)$$

Remark 1 Note that these results are coherent with that of the Markovian case (exponential service time distributions) obtained in the work [11]. More precisely, we showed that formulas (24), (25) and (26) are valid in the case of non-exponential distributions of service times for both incoming and outgoing calls.

7. MEAN PERFORMANCE METRICS

7.1 First moments of the queue lengths

We provide now some mean performance measures of interest. Let $N = \lim_{t\to\infty} N(t)$ denote the number of customers in the orbit in the steady state.

1. Mean number of customers in orbit when server is idle:
$$M_1^0 = E(z^N, S=0)|_{z=1} = Q_0'(1,\infty) =$$
$$= \frac{2\lambda(\rho+\sigma)(1-\rho) + \lambda^3(1+\sigma)h_{12} + (1-\rho)\lambda^2\alpha h_{22}}{2[\mu(1-\rho)-\lambda(\rho+\sigma)](1+\sigma)}. \quad (22)$$
2. Mean number of customers in orbit when the server is busy by the service of incoming call:
$$M_1^1 = E(z^N, S=1)|_{z=1} = Q_1'(1,\infty) =$$
$$(\lambda+\mu)h_1 Q_0'(1) - \frac{\rho(\rho+\sigma)}{1+\sigma} + \frac{\lambda^2}{2}h_{12}. \quad (23)$$
3. Mean number of customers in orbit when the server is busy by the service of outgoing call:
$$M_1^2 = E(z^N, S=2)|_{z=1} = Q_2'(1,\infty) =$$
$$= \alpha h_2 Q_0'(1,\infty) + \frac{\alpha\lambda h_{22}}{2}Q_0(1,\infty). \quad (24)$$
4. Unconditional mean number of customers in orbit:
$$E[N] = Q'(z) = E(z^N)|_{z=1} = M_1^0 + M_1^1 + M_1^2. \quad (25)$$

7.2 Second moments of the queue length

Next we calculate the second partial factorial moments of the joint queue length.

$$M_2^i = Q_i(z)''|_{z=1}, \qquad i = 0, 1, 2.$$

We have obtained explicit forms for the partial generating functions $Q_i(z)$ $(i = 0, 1, 2)$. Thus, in principle, factorial moments can be obtained by differentiating these generating functions and applying L'Hospital's rule at $z = 1$. However, the computation is rather complicated. Here we apply the Taylor series expansion method presented in [12] providing a systematic way to obtain all the factorial moments of the joint queue lengths.

We have

$$h_1(\lambda - \lambda z) = 1 + \rho(z-1) + \sum_{n=2}^{\infty} \frac{\lambda^n h_{1n}}{n!}(z-1)^n, \qquad (26)$$

$$h_2(\lambda - \lambda z) = 1 + \sigma(z-1) + \sum_{n=2}^{\infty} \frac{\lambda^n h_{2n}}{n!}(z-1)^n, \qquad (27)$$

where $\rho = \lambda h_1$ and $\sigma = \lambda h_2$. Furthermore, h_{1n} and h_{2n} are n-th moments of incoming and outgoing calls.

First, we rewrite $Q_0(z)$ as follows.

$$Q_0(z) = \frac{\mu z - \mu h_1(\lambda - \lambda z)}{(\lambda+\mu)z - (\lambda z + \mu)h_1(\lambda - \lambda z) + \alpha z(1 - h_2(\lambda - \lambda z))} P_{00}.$$

It is easy to check that both the nominator and denominator of the right hand side of the above equation vanish at $z = 1$. Using Taylor series expansion at $z = 1$, i.e., formulae (26) and (27) and dividing both the nominator and denominator by $(z-1)$ yields,

$$Q_0(z) = \frac{Q_0^{(n)} + Q_1^{(n)}(z-1) + Q_2^{(n)}(z-1)^2 + O((z-1)^2)}{Q_0^{(d)} + Q_1^{(d)}(z-1) + Q_2^{(d)}(z-1)^2 + O((z-1)^2)} P_{00}, \qquad (28)$$

where $O((z-1)^n)$ includes $(z-1)^{(n+1)}$ and higher order terms. Furthermore,

$$Q_0^{(n)} = \mu(1-\rho), \; Q_1^{(n)} = -\mu\frac{\lambda^2 h_{12}}{2}, \; Q_2^{(n)} = -\mu\frac{\lambda^3 h_{13}}{3!}.$$

and

$$Q_0^{(d)} = \mu - \rho(\lambda+\mu) - \lambda\sigma,$$

$$Q_1^{(d)} = -\left(\lambda\rho + (\lambda+\mu)\frac{\lambda^2 h_{12}}{2} + \alpha\frac{\lambda^2 h_{22}}{2} + \lambda\sigma\right),$$

$$Q_2^{(d)} = -\left(\frac{\lambda^3 h_{12}}{2} + (\lambda+\mu)\frac{\lambda^3 h_{13}}{3!} + \frac{\alpha\lambda^2 h_{22}}{2} + \frac{\alpha\lambda^3 h_{23}}{3!}\right).$$

We have

$$Q_0(z) = M_0^0 + M_1^0(z-1) + M_2^0\frac{(z-1)^2}{2} + O((z-1)^2). \quad (29)$$

Transforming (28) as follows.

$$Q_0(z)\left(Q_0^{(d)} + Q_1^{(d)}(z-1) + Q_2^{(d)}(z-1)^2 + O((z-1)^2)\right)$$
$$= Q_0^{(n)} + Q_1^{(n)}(z-1) + Q_2^{(n)}(z-1)^2 + O((z-1)^2).$$

Substituting (29) into this equation and comparing the coefficients of $(z-1)^2$ in both sides yields

$$\frac{M_2^0 Q_0^{(d)}}{2} = P_{00}Q_2^{(n)} - M_1^0 Q_1^{(d)} - M_0^0 Q_2^{(d)},$$

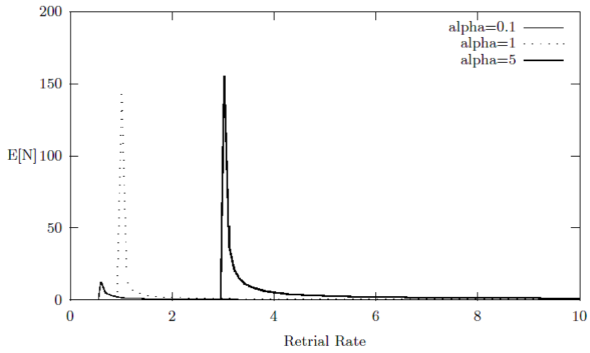

Figure 1: Mean number of calls in the queue $E[N]$ versus retrial rate μ when varying outgoing rate α.

which is equivalent to

$$M_2^0 = \frac{2(P_{00}Q_2^{(n)} - M_1^0 Q_1^{(d)} - M_0^0 Q_2^{(d)})}{Q_0^{(d)}}.$$

Furthermore, comparing the coefficients of $(z-1)^3$, we obtain an expression of M_3^0 in terms of M_2^0. So, we can calculate all the factorial moments M_n^0 $(n \geq 3)$ in a recursive manner.

Similarly, comparing the coefficients of $(z-1)^2$ in the expressions for $Q_1(z, \infty)$, we have

$$M_2^1 = \left(2\lambda M_1^0 + (\lambda+\mu)M_2^0\right)h_1$$
$$+\lambda h_{12}\left(\lambda M_0^0 + (\lambda+\mu)M_1^0\right)\big)$$
$$+\frac{\lambda^2 h_{13}}{3}\left((\lambda+\mu)M_0^0 - \mu P_{00}\right) - 2M_1^1.$$

Finally, also comparing the coefficients of $(z-1)^2$ in the expression of $Q_2(z, \infty)$ yields

$$M_2^2 = \alpha\left(h_{21}M_2^0 + \lambda h_{22}M_1^0 + \frac{\lambda^2 h_{23}}{3}M_0^0\right).$$

8. NUMERICAL ILLUSTRATIONS AND EXAMPLES

In this section, we give some examples of how managers can exploit the obtained results. In all cases, we take exponential service time probability distributions, without loss of generality, since the first moments depend only on the first two moments of (nonexponential) parametric distributions (service times for both incoming and outgoing calls). Incoming calls arrive according to a Poisson process with rate $\lambda = 0.5$.

8.1 Effect of retrial and outgoing rates on the mean number of calls in the queue

Figure 1 shows the evolution of the mean number of customers in orbit in stationary regime $E[N]$ in function of the retrial parameter μ when varying the outgoing rate $\alpha = 0.1$, $\alpha = 1$ and $\alpha = 5$. We fixed some parameters as follows: $\lambda = 0.5; h_1 = 1; h_2 = 0.5; h_{12} = 0.5; h_{22} = 1$. We see that $E[N]$ increases when the outgoing call rate α increases, but tends rapidly to zero when the retrial rate increases.

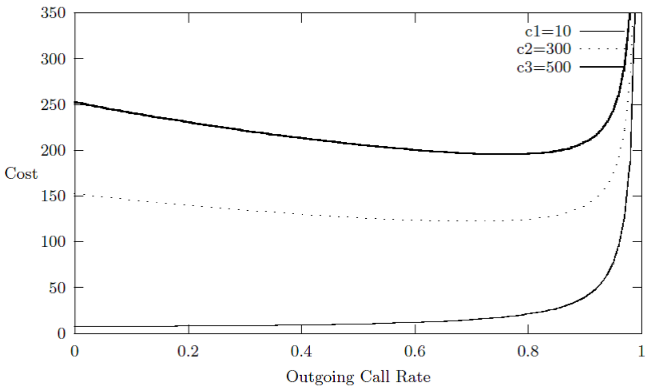

Figure 2: Cost as a function of varying the outgoing call rate α with various c_1.

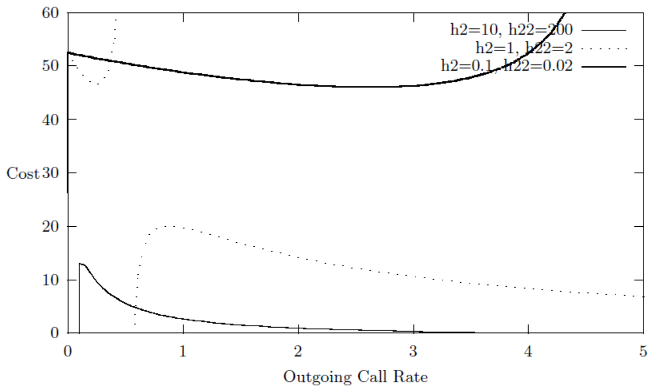

Figure 3: Cost as a function of varying the outgoing call rate α with various h_2.

8.2 Optimisation of Idle time

There is indeed several optimisation problems which can be associated to such questions. But, we restrict ourself to the one's introduced in [11] which concerns our subject of the profitability of idle time of the server. The derivation of performance measures of Section 6 and Section 7 allow to formulate the following cost model. Let U denote the utilization of the server defined by $U = M_0^1 + M_0^2$. One operational objective in the context of call center is to show the optimal value of α which minimizes simultaneously $1-U$ and the mean number of customers in orbit $E[N]$, i.e.,

$$C(\alpha) = c_1(1-U) + c_2 E[N],$$

under the constraint

$$\frac{\lambda+\mu}{\mu}\rho + \frac{\lambda}{\mu}\sigma < 1$$

where c_1 is the cost of idle server and c_2 the cost of a retrial customer. The constraint inequality is just the stability condition. The motivation for this cost formulation is that the value of α can be controlled by the operator directly, or by the ACD automatically.

Table 1: Parameter setting for Figs. 2 and 3.

Figure	λ	h_1	h_2	μ	c_1	c_2
2	0.5	1	2	1	*Various*	1
3	0.5	1	*Various*	1	100	1

Consider the case of exponential distribution of service for outgoing calls. Since this model concerns the case with general distributions of service, the expression for $E[N]$ depends not only on the service rate $\nu_2 = \frac{1}{h_2}$ (in the notations of ([11]), but also on the second moments h_{12} and h_{22}. So, we need to fix these parameters according to the exponential case. For example, since $h_1 = \nu_1 = 1$, we must have $h_{12} = \frac{2}{\nu_1^2} = 2$. Similarly, $h_{22} = 200$ when $h_2 = 10$ ($\nu_2 = 0.1$), $h_{22} = 2$ when $h_2 = 1$ ($\nu_2 = 1$) and $h_{22} = 0.02$ when $h_2 = 0.1$ ($\nu_2 = 10$).

The curves of Fig. 2 are the same of Fig. 1(a) in [11]. So, the same discussion yields. We observe some difference when varying the mean service rate of outgoing calls h_2 (or its rate ν_2). For various h_2, illustrated in Fig. 3, the cost function seems to have non analytical regular structure due to the constraint. For $h_2 = 1$ (the curve in dot line), the admissible values of α lies on the interval $(0; 0.5)$. So, the optimal value $\alpha_{opt} = 0.5$. For $h_2 = 10$ (the lower continue line), the admissible values of α lies on the interval $(0; 0.05)$. So, the optimal value $\alpha_{opt} = 0$. Finally, when $h_2 = 0.1$ (the upper continue line), the eligible values of α lies on the interval $(0; 5)$. So, the optimal value of α_{opt} is visibly in the neighbour of $\alpha = 3$. So, it seems that the optimal value α_{opt} decreases with decreasing of h_2 which is coherent with the conclusions of [11].

9. CONCLUSION

We have provided some advances in the model $M/G/1$ with two-way communication system with constant retrial policy and general service times. In perspective, we aim to study the case of general constant retrials. Indeed, in this case the mean characteristics will depend on the probability distribution of retrial times (in opposite to the case of linear retrial rate). More insight needs to be provided into optimisation problems in relation with the exploitation of idle time.

10. REFERENCES

[1] A. Aissani. An $MX/G/1$ energetic retrial queue with vacations and control. IMA Journal of Management Mathematics, 22:13–32, 2011.

[2] A. Aissani and T.Phung-Duc.Optimal analysis for $M/G/1$ retrial queue with two-way communication. Analytical and Stochastic Modelling Techniques and Applications, Lecture Notes in Computer Science, 9081:1–14, 2015.

[3] J.R. Artalejo. Analysis of an $M/G/1$ queue with constant repeated attempts and server vacations, Computers Operations Research, 24:493–504, 1997.

[4] J.R. Artalejo, A. Gomez-Corral. Retrial Queueing Systems: A Computational Approach.Springer, Berlin, 2008.

[5] J.R. Artalejo, T. Phung-Duc. Markovian single server retrial queues with two way communication. Proceedings

of the 6th International Conference on Queuieng Theory and Network Applications, QTNA'11, 1–7,ACM, New York, 2011.

[6] J.R. Artalejo and T. Phung-Duc.,Markovian retrial queues with two way communication. Journal of Industrial and Management Optimization, 8:781-806, 2012.

[7] J.R. Artalejo and T. Phung-Duc. Single server retrial queues with two way communication. Applied Mathematical Modelling, 37:1811-1822, 2013.

[8] B.V. Gnedenko and I.N. Kovalenko. Introduction to queueing theory, Birkhauser, Boston, 1989.

[9] G. Koole and A. Mandelbaum.Queueing models of call centers: an introduction. Annals of Operations Research,113:41–59,2002.

[10] G.I. Falin, J.G.C. Templeton. Retrial Queues, Chapman and Hall, London, 1997.

[11] T. Phung-Duc, W. Rogiest, Y. Takahashi and H. Bruneel. Retrial queues with balanced call blending: analysis of single-server and multi-server queue. Annals of Operations Research, 239:429-449, 2016.

[12] T. Phung-Duc. Single server retrial queues with setup time. Journal of Industrial and Management Optimization, 13:1329–1345, 2017.

[13] H. Sakurai and T. Phung-Duc. Two-way communication retrial queues with multiple types of outgoing calls, TOP, 23:466-492, 2015.

[14] H. Sakurai and T. Phung-Duc. Scaling limits for single server retrial queues with two-way communication. Annals of Operations Research, 247:229-256, 2016.