Blended call center with idling times during the call service

Benjamin Legros, Oualid Jouini, Ger Koole

Blended Call Center with Idling Times during the Call Service

Benjamin Legros[a] • Oualid Jouini[b] • Ger Koole[c]

[a] *EM Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016, Paris, France*

[b] *CentraleSupélec, Université Paris-Saclay, Laboratoire Genie Industriel, Grande Voie des Vignes, 92290 Chatenay-Malabry, France*

[c] *VU University Amsterdam, Department of Mathematics, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*

benjamin.legros@centraliens.net • oualid.jouini@centralesupelec.fr • ger.koole@vu.nl

Abstract

We consider a blended call center with calls arriving over time and an infinitely backlogged amount of outbound jobs. Inbound calls have a non-preemptive priority over outbound jobs. The inbound call service is characterized by three successive stages where the second one is a break, i.e., there is no required interaction between the customer and the agent for a non-negligible duration. This leads to a new opportunity to efficiently split the agent time between inbound calls and outbound jobs.

We focus on the optimization of the outbound job routing to agents. The objective is to maximize the expected throughput of outbound jobs subject to a constraint on the inbound call waiting time. We develop a general framework with two parameters for the outbound job routing to agents. One parameter controls the routing between calls, and the other does the control inside a call. We then derive structural results with regard to the optimization problem and numerically illustrate them. Various guidelines to call center managers are provided. In particular, we prove for the optimal routing that at least one of the two outbound job routing parameters has an extreme value.

Keywords: Call centers; blending; task overlapping; routing; performance evaluation; waiting times; optimization; queueing systems; Markov chains.

1 Introduction

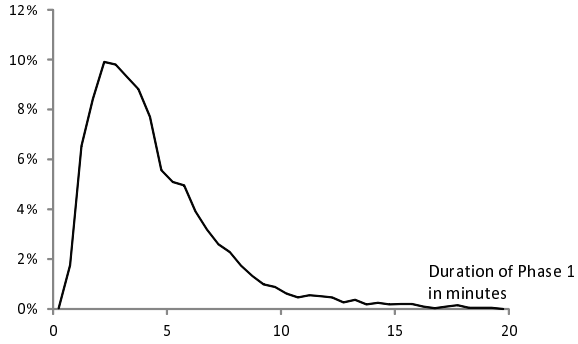
Context and Motivation: New technology-driven innovations in call centers are multiplying the opportunities to make more efficient use of an agent as she can handle different types of workflow, including inbound calls, outbound calls, emails and chats. However, several issues on the management of call center operations emerged also as a result of advanced technology. In this paper, we consider a call center with two types of jobs, inbound and outbound jobs. We focus on how to efficiently share the agent time between the two types of jobs in order to improve the call center performance.

Call center situations where inbound calls and outbound jobs (outbound calls or emails) are combined is referred to as *blending*. The key distinction of problems with blending comes from the fact that outbound jobs are less urgent and can be inventoried to some extent, relative to incoming calls. Therefore, managers are likely to give a strict priority to inbound calls over outbound jobs. An important question here is what should be the best way of routing outbound jobs to agents, i.e., as a function of the system parameters and the service level constraints (on calls and outbound jobs), when should we ask the agent to treat outbound jobs between the call conversations (Bernett et al., 2002; Bhulai and Koole, 2003; Pang and Perry, 2014; Legros et al., 2015). The outbound job routing question is further important in the context of the call center applications we consider here. We encountered examples where a call conversation between an agent and a customer contains a *natural break*. We mean by this a time interval with no interaction between the agent and the customer. During the conversation, the agent asks the customer to do some necessary operations in her own (without the need of the agent availability). After finishing those operations, the conversation between the two parties can start again. Inside an underway conversation, the agent is then free to do another job if needed.

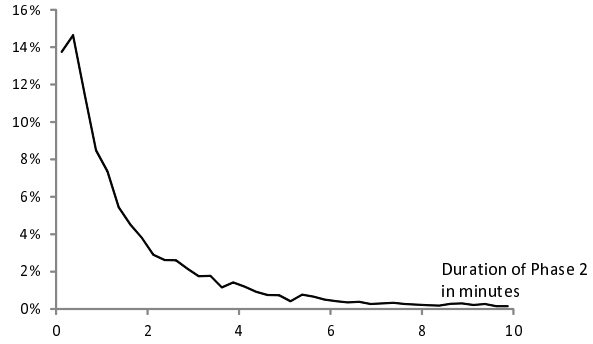
For an efficient use of the agent time, one would think about the routing of the less urgent jobs not only when the system is empty of calls, but also during call conversations. In practice, such a situation often occurs. For example, an agent in an internet hotline call center asks the customer to reboot her modem or her computer which may take some time where no interactions can take place. It is also often the case that a call center agent of an electricity supplier company asks the customer for the serial number of her electricity meter box. This box is usually located outside of the house and is locked, so, the customer needs some non-negligible time to get the required information. Another example is that of commercial call centers with a financial transaction during the call conversation. After some time from the start of the call conversation, the customer is asked to do an online payment on a website before coming back to the same agent in order to finish the conversation. The online payment requires the customer to locate her credit card, next she enters the credit card numbers, next she goes through the automated safety check with her bank (using SMS for example), which may take some minutes.

As an illustration, we provide real-life data from a vehicle glass repair call center company where the service process consists of three phases and a break in the second phase. Figure 1 gives the empirical probability density functions of the three service phase durations for 5986 calls. We observe that these durations are random and find a good log-normal fit. The $p-$values associated with the statistical χ^2 tests are all above the threshold of 5%. For phases 1, 2 and 3, they are 0.056, 0.561 and 0.101, respectively. In particular, we observe that the average break duration (phase 2) is 2.39 min, and represents in average around 30% of the total service duration. It is natural for such a setting that the system manager thinks about using the opportunity to route outbound jobs

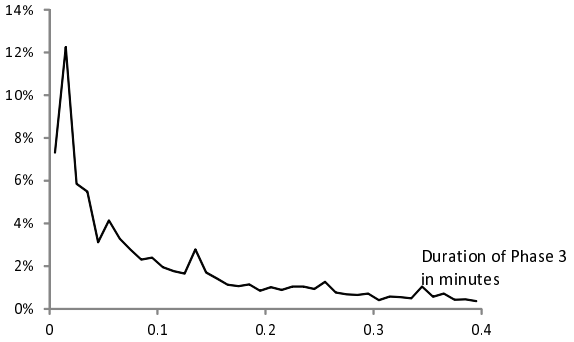
(or back-office tasks in general) to an agent during the break of an ongoing call conversation, and not only when no calls are waiting in the queue. The advantage is an efficient use of the agent time and therefore a better call center performance.



(a) Service phase 1



(b) Service phase 2



(c) Service phase 3

Figure 1: Empirical probability density functions of phase durations

Main Contributions: We consider a call center with an infinite amount of outbound jobs. Inbound calls arrive over time, and in the middle of an inbound call conversation, a break is required. Given this type of call centers, we are interested in optimizing its functioning by controlling how the resource should be shared between the two types of jobs. Calls are more important than outbound jobs in the sense that calls request a quasi-instantaneous answer (waiting time in the order of some minutes), however outbound jobs are more flexible and could be delayed for several hours. An appropriate functioning is therefore that an agent works on inbound calls as long as there is work to do for them. The agent can then work on outbound jobs when she becomes free from calls, i.e., after a service completion when no calls are waiting in the queue, or during the call conversation break. We assume that calls have a strict non-preemptive priority over outbound jobs, which means that if a call is busy with an outbound job (that has started after a service completion or during the break), the agent will finish first the outbound job before turning to a new arrived call to the queue or a call that has accomplished the requested operations and wants to start again the conversation to finish her service. The non-preemption priority rule is coherent with the operations in practice

and also to the call center literature (Bhulai and Koole, 2003; Deslauriers et al., 2007). It is not appropriate to stop the service of a low priority customer.

We focus on the research question: when should the agent treat outbound jobs? Between calls, or inside a call conversation, or in both situations? Given the nature of the job types, a call center manager in practice would be interested in maximizing the number of treated outbound jobs while respecting some service level objective on the call waiting time (Bhulai and Koole, 2003). For inbound calls, we are interested in the steady-state performance measures in terms of the expected waiting time, the probability that the waiting time is less than a given threshold, and the probability of delay. For outbound jobs, we are interested in the steady-state performance in terms of the expected throughput, i.e., the number of treated outbound jobs per unit of time.

Despite its prevalence in practice, there are no papers in the call center literature addressing such a question. Most of the related papers only focus on the outbound job routing between call conversations but not inside a call conversation. To answer this question, we develop a general framework with two parameters for the outbound job routing to agents. One parameter controls the routing between calls, and the other does the control inside a call conversation. Although this modeling is not optimal, its performance measures as shown later are close to the optimal ones and its routing policy is easier to implement than the complex optimal routing. For the tractability of the analysis, we first focus on the single server case. We then discuss the extension of the results to the multi-server case and its applicability to a more complex setting with may include abandonment, general service time distribution and a time-dependent arrival process. For the single server modeling, we first evaluate the performance measures using a Markov chain analysis. Second, we propose an optimization method of the routing parameters for the problem of maximizing the outbound job expected throughput under a constraint on the service level of the call waiting time. As a function of the system parameters (the server utilization, the outbound job service time, the severity of the call service level constraint, etc.), we derive various guidelines to managers. In particular, we prove for the optimal routing that at least one of the two outbound job routing parameters has an extreme value. As detailed later, an extreme value means that the agent should always do outbound jobs inside a call (or between calls) or not at all. In other cases, the parameters lead to randomized policies. We also solve the optimization problem by proposing four particular cases corresponding to the extreme values of the probabilistic parameters. We analytically derive the conditions under which one particular case would be preferred to another one. The interest from the particular cases is that they are easy to understand by agents and managers. Several numerical experiments are used to illustrate the analysis. We then focus on the routing optimization problem for the multi-server case in a more general setting, using simulation and approximations developed under the light and heavy-traffic regimes. We found that most of the observations of the single server case are still valid (in particular the result stating that at least one control parameter has

an extreme value). This justifies the applicability of our results to real-life call centers.

Paper Organization: The rest of the paper is organized as follows. In Section 2 we review some of the related literature. In Section 3, we describe the blended call center modeling and the optimization problem. In Section 4, we consider a single server analysis and develop a method based on the analysis of Markov chains in order to derive the performance measures of interest for inbound and outbound jobs. Next, we focus on optimizing the outbound job routing parameters. In Section 5, we extend the previous analysis to the multi-server case using simulation and also asymptotic approximations. In Section 6, we assess the applicability of our analysis to a more setup with abandonment, general service time distribution and time-dependent arrival process. Finally, Section 7 concludes the article.

2 Literature Review

There are three related streams of literature to this paper. The first one deals with blended call centers. The second one is the Markov chain analysis for queueing systems with phase-type service time distributions. The third one is related to the cognitive analysis, or in other words the ability for an agent to treat and switch between different job types.

The literature on blended call centers consists of developing performance evaluation and optimal blending policies. Deslauriers et al. (2007) develop a Markov chain for the modeling of a Bell Canada blended call center with inbound and outbound calls. The performance measures of interest are the rate of outbound calls and the waiting time of inbound calls. Through simulation experiments they prove the efficiency of their Markov chain model to reflect reality. Brandt and Brandt (1999) develop an approximation method to evaluate the performance of a call center model with impatient inbound calls and infinitely patient outbound calls of lower priority than the inbound traffic. Bhulai and Koole (2003) consider a similar model to the one analyzed in this paper, except that the call service is done in a single stage without a possible break. The model consists of inbound and outbound jobs where the inbound jobs have a non-preemptive priority over the outbound ones. For the special case of identically distributed service times for the two jobs, they optimize the outbound jobs routing subject to a constraint on the expected waiting time of inbound jobs. Gans and Zhou (2003) study a call center with two job types where one of the jobs is an infinitely backlogged queue. They develop a routing policy consisting in the reservation of servers in order to maximize the expected throughput on the jobs of the infinitely backlogged queue. Armony and Maglaras (2004) and Legros et al. (2016) analyze a similar model with a call-back option for incoming customers. The customer behavior is captured through a probabilistic choice model. Other references include (Bernett et al., 2002; Keblis and Chen, 2006; Pichitlamken et al., 2003).

The analysis in this paper is also related to the analysis of queueing systems with phase-type

service time distributions. We model the call service time through three successive exponentially distributed stages, where the second stage may also overlap with the service of one or several outbound jobs with an exponential time duration for each. The performance evaluation of such systems involves the steady-state analysis of Markov chains and is usually addressed using numerical methods. We refer the reader to Kleinrock (1975) for simple models with Erlang service time distributions. For more complex systems, the reader is refereed to Bolotin (1994); Brown et al. (2005); Guo and Zipkin (2008). Our approach to derive the performance measures is based on first deriving the stationary system state probabilities for two-dimension and semi-infinite continuous time Markov chains. One may find in the literature three methods for solving such models. The first one is to truncate the state space, see for example Seelen (1986) and Keilson et al. (1987). The second method is called spectral expansion (Daigle and Lucantoni, 1991; Mitrani and Chakka, 1995; Choudhury et al., 1995). It is based on expressing the invariant vector of the process in terms of the eigenvalues and the eigenvectors of a matrix polynomial. The third one is the matrix-geometric method (Neuts, 1995). The approach relies on determining the minimal positive solution of a non-linear matrix equation. The invariant vector is then expressed in terms of powers of itself. In our analysis, we reduce the problem to solving cubic and quartic equations, for which we use the method of Cardan and Ferrari (Gourdon, 1994).

Finally, we briefly mention some studies on human multi-tasking, as it is the case for the agents in our setting. Gladstones et al. (1989) show that a simultaneous treatment of jobs is not efficient even with two easy jobs because of the possible interferences. In our models, we are not considering simultaneous tasks in the sense that an agent cannot talk to a customer and at the same time treats an outbound job. More interestingly, Charron and Koechlin (2010) studied the capacity of the frontal lobe to deal with different tasks by alternation (as here for calls and outbound jobs). They develop the notion of *branching*: capacity of the brain to remember information while doing something else. They show that the number of jobs done alternatively has to be limited to two to avoid loss of information. Dux et al. (2009) showed that training and experience can improve multi-tasking performance. The risk from alternating between two tasks is the loss of efficiency because of switching times. An important aspect to avoid inefficiency as pointed out by Dux et al. (2009) and Charron and Koechlin (2010) is that the alternation should be at most between two tasks quite different in nature (like inbound and outbound jobs).

3 Problem Description and Modeling

We consider a call center modeling with s identical agents and two types of jobs: inbound calls and outbound jobs. The arrival process of inbound calls is assumed to be Poisson with mean arrival rate λ. There is an infinite amount of outbound jobs (emails, back-office tasks, etc.) that are waiting

to be treated in a dedicated first come, first served (FCFS) queue with infinite capacity.

We consider call center applications where the communication between the agent and the customer includes a break (the customer does not need the agent availability). We model the service time of a call by three successive stages. The first stage is a conversation between the two parties. The second stage is the break, i.e., no interactions between the two parties. The third and final step is again a conversation between the two parties. The service completion occurs as soon as the third stage finishes. We model each stage duration as an exponentially distributed random variable, with rate μ_i for stage i. The durations of the three stages are jointly independent. This Markovian assumption, which is common in modeling in service operations, is reasonable for systems with high service time variability where service times are typically small but there are occasionally long service times. An agent handles an outbound job within one single step without interruption. The time duration of an outbound job treatment is random and assumed to be exponentially distributed with rate μ_0. Moreover, we assume the service durations are not known by the system manager before realization. So, the routing decisions cannot be based on such information. This is a common assumption for call centers. For other applications, like packet delivery on internet, the random variables may be realized before service. Examples of related studies include Zhang (1995); Le Boudec (1998); Gevros et al. (2001). The queueing model is depicted in Figure 2.

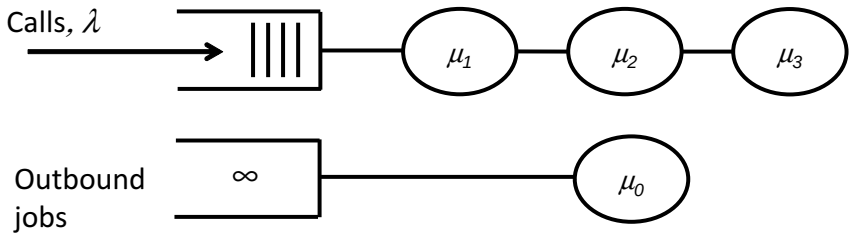


Figure 2: Queueing model

Bhulai and Koole (2003) and Gans and Zhou (2003) consider a similar model with inbound customers who arrive according to a homogeneous Poisson process and an infinite amount of outbound jobs, exactly like in our model. Their objective is to maximize the throughput of served outbounds with a constraint on the waiting time of inbounds. They show, using a Markov decision process approach, that a strict priority should be given for inbound calls. Using their results, we also assume the strict priority of inbound calls. More precisely, upon arrival, a call is immediately handled by an available agent, if any. If not, the call waits for service in an infinite FCFS dedicated queue. Inbound calls have a non-preemptive priority over outbound jobs. Non-preemption is a natural assumption for our application since outbound jobs could be for example outbound calls. We are interested in an efficient use of the agent time between inbound calls and outbound jobs. More concretely, we want to answer the question when should we treat outbound jobs for the

following optimization problem

$$\begin{cases} \text{Maximize the expected throughput of outbound jobs} \\ \text{subject to a service level constraint on the call waiting time in the queue.} \end{cases} \quad (1)$$

We numerically address Problem (1) using a Markov Decision Process (MDP) approach. As shown in Section 1 of the online appendix, the optimal policy is very complex. It depends on six parameters: the number of inbound calls waiting in the queue, the number of agents working on each stage of service, the number of outbound jobs being in service between calls and the number of outbound jobs being in service inside the break. This makes the optimal policy hard to obtain in general and then hard to implement in practice. We instead propose a simpler model for the routing of outbound jobs to agents. It is referred to as *probabilistic model* or *Model PM* and is described below. Although this model is not optimal, we numerically show its efficiency through a comparison with the optimal policy (see Section 1 of the online appendix). It is moreover easy to implement and to understand by a system manager.

Probabilistic Model (Model PM): We distinguish the two situations when an agent is available to handle outbound jobs between two call conversations, or inside a call conversation.
Between two calls: just after a call service completion (as soon as the third stage finishes) and no waiting calls are in the queue, the agent treats one or more outbound jobs with probability p (independently of any other event), or does not work on outbound jobs at all with probability $1-p$. In the latter case, the agent simply remains idle and waits for a new call arrival to handle it. In the former case (with probability p), she selects a first outbound job to work on. After finishing the treatment of this outbound job, there are two cases: either a new call has already arrived and it is now waiting in the queue, or the queue of calls is still empty. If a call has arrived, the agent handles that call. If not, she selects another outbound job, and so on. At some point in time, a new call would arrive while the agent is working on an outbound job. The agent will then handle the call as soon as she finishes the outbound job treatment.
Inside a call: Just after the end of the first stage of an underway call service (regardless whether there are other waiting calls in the queue or not), the agent treats one or more outbound jobs with probability q (independently of any other event), or does not work on outbound jobs at all with probability $1-q$. In the latter case, the agent simply remains idle and waits for the currently served customer to finish her operations on her own (corresponding to the second call service stage, i.e., the agent break). As soon as the customer finishes by herself her second service stage, the agent starts the third and last service stage. In the former case (with probability q), she selects a first outbound job to work on. After finishing the treatment of this outbound job, there are two cases: either the currently served customer has already finished her second service stage, or not yet. If

Table 1: Particular cases of Model PM

Model	Description
Model 1	$p = q = 0$, no treatment of outbound jobs
Model 2	$p = 1$ and $q = 0$, systematic treatment of outbound jobs only between two calls
Model 3	$p = 0$ and $q = 1$, systematic treatment of outbound jobs only during the break
Model 4	$p = q = 1$, systematic treatment of outbound jobs between two calls and during the break

she does, the agent starts the third stage of the customer call service. If not, she selects another outbound job, and so on. At some point in time, the currently served call would finish her second service while the agent is working on an outbound job. The agent will then handle the call as soon as she finishes the the outbound job treatment. Model PM is depicted in Figure 3.

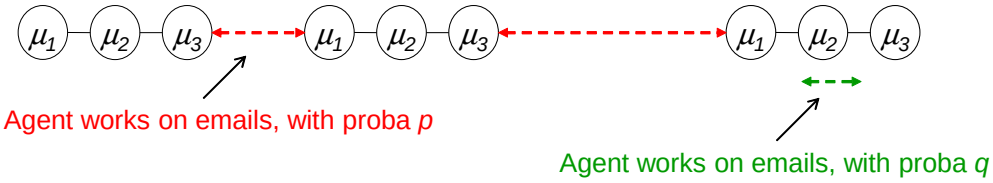


Figure 3: Model PM

The decision to initiate an outbound job is only taken at an inbound service completion (control parameter p) or at the completion of the first phase of service of an inbound (control parameter q). After each outbound job service completion, a new outbound service is automatically initiated. One may think of modifying Model PM by allowing to switch and stop serving outbound jobs. Actually, due to the exponential assumptions for the inter-arrival, the second phase of service and the outbound service times, Model PM is equivalent to a model where a new decision may be taken upon each outbound service completion. The proof of the result is given in Section 2 of the online appendix.

We further consider next four particular cases of Model PM as shown in Table 1. Although these models might appear to be too restrictive to solve Problem (1), we show later their merit in Section 4.2.2 when we focus on the optimization of p and q in Model PM. Moreover, they have the advantage of being easy to implement in practice, easy to understand by managers, and easy to follow by agents. Note that in Model 1, the expected throughput of outbound jobs is zero. The interest from Model 1 is in the extreme case of a very high workload of calls or a very restrictive constraint on the call waiting time.

The objective for the system manager is to find the optimal values for the control parameters p and q or to determine among the particular cases of Model PM which one would better answer Problem (1). The knowledge of the system parameters is essential to obtain implementable results. Section 3 of the online appendix is devoted to the estimation of these parameters based on real

data.

4 Single Server Analysis

In this section, we provide an exact method to characterize the call waiting time in the queue and the outbound job expected throughput for Model PM and its extreme cases for a single server model. We also develop various structural results for the optimization problem. The tractable analysis in this section enhances our understanding of the system behavior. This would not be possible to do directly for the multi-server case since an exact analysis is very complex. However, we extend the analysis to multi-server case in Section 5 using light and heavy-traffic approximations.

Our approach consists of using a Markov chain model to describe the system states and compute their steady-state probabilities. The computation of some of the steady-state probabilities involves the resolution of cubic (third degree) or quartic (fourth degree) equations for which we use the Cardan-Ferrari method.

4.1 Performance Evaluation

Let us define the random process $\{(x(t), y(t)), t \geq 0\}$ where $x(t)$ and $y(t)$ denote the state of the agent and the number of waiting calls in the queue at a given time $t \geq 0$, respectively. We have $y(t) \in \{0, 1, 2, ...\}$, for $t \geq 0$. The possible values of $x(t)$ (corresponding to the possible states of the agent), for $t \geq 0$, are

- "Agent working on the first stage of a call service" denoted by $x(t) = A$,
- "Idle agent that is waiting for the call to finish her second stage of service" denoted by $x(t) = B$,
- "Agent working on an outbound job while an underway call has already finished her second stage of service and is waiting for the agent to start her third stage of service" denoted by $x(t) = B'$,
- "Agent working on the third stage of a call service" denoted by $x(t) = C$,
- "Agent working on an outbound job between two call conversations" denoted by $x(t) = M$,
- "Agent idle between two call conversations" denoted by $x(t) = 0$.

Since call inter-arrival times, call service times in each stage, and outbound job service times are exponentially distributed, $\{(x(t), y(t)), t \geq 0\}$ is a Markov chain (Figure 4).

For ease of exposition, we denote by P_0 the probability to be in state $(0, 0)$, and for $n \geq 0$ we denote by a_n, b_n, b'_n, c_n and m_n the probabilities to be in state (A, n), (B, n), (B', n), (C, n) and (M, n), respectively. We also define $\rho_i = \frac{\lambda}{\mu_i}$, for $i \in \{0, 1, 2, 3\}$. In Proposition 1, we give the probability of delay of a call (probability of waiting) denoted by P_D and the expected throughput of outbound jobs denoted by T. Note that the stability condition of Model PM is $\lambda < \frac{1}{\frac{q}{\mu_0}+\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}}$.

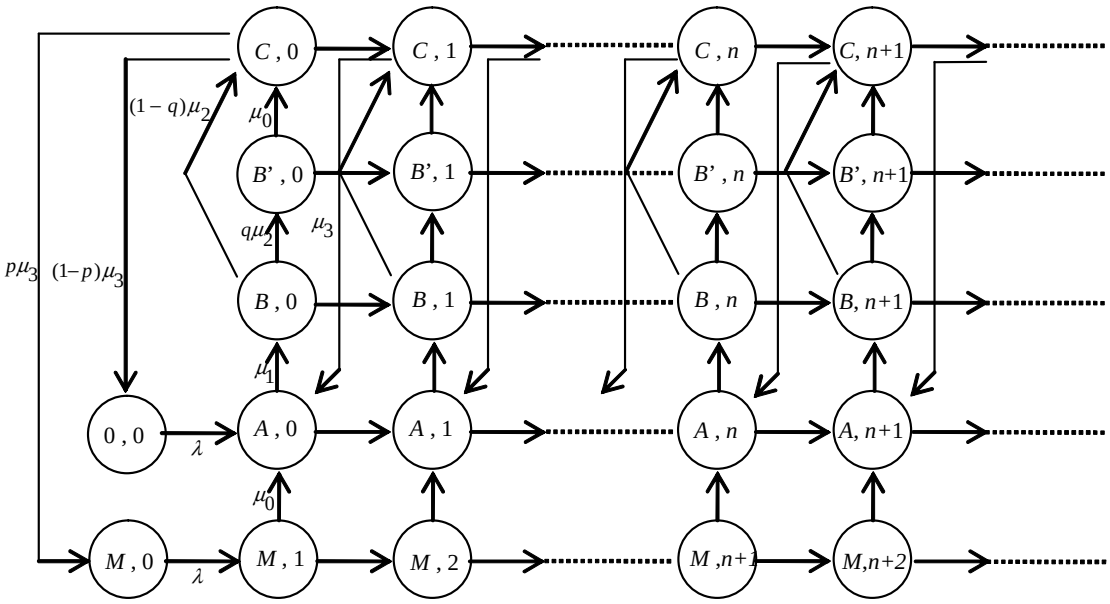


Figure 4: Markov chain for Model PM

Proposition 1 *For Model PM, we have*

$$P_D = 1 - \frac{1-p}{1+p\rho_0}(1-\rho_1-\rho_2-q\rho_0-\rho_3),$$
$$T = \mu_0\left(\frac{1+\rho_0}{1+p\rho_0}p\,(1-\rho_1-\rho_2-q\rho_0-\rho_3)+q(\rho_2+\rho_0)\right).$$

Proof. From the Markov chain of Model PM, we have

$$c_0 = \rho_3(P_0+m_0),$$
$$c_n = \rho_3(a_{n-1}+b_{n-1}+b'_{n-1}+c_{n-1}+m_n),$$

for $n \geq 1$. Then

$$\sum_{n=0}^{\infty} c_n = \rho_3\left\{P_0+m_0+\sum_{n=0}^{\infty}\left(a_n+b_n+b'_n+c_n\right)+\sum_{n=1}^{\infty}m_n\right\}. \tag{2}$$

Since all system state probabilities sum up to 1, i.e., $P_0+\sum_{n=0}^{\infty}(a_n+b_n+b'_n+c_n+m_n)=1$, Equation (2) becomes

$$\sum_{n=0}^{\infty} c_n = \rho_3. \tag{3}$$

For the state $(M,0)$, we have $p\mu_3 c_0 = \lambda m_0$, or equivalently $c_0 = \rho_3\frac{m_0}{p}$. Therefore $c_0 = \rho_3\frac{P_0}{1-p}$. We

then may write

$$P_0 = \frac{1-p}{p} m_0. \tag{4}$$

Combining Equation (3) and the relations $\mu_2 \sum_{n=0}^{\infty} b_n = \mu_3 \sum_{n=0}^{\infty} c_n = \mu_0 \sum_{n=0}^{\infty} b'_n + (1-q)\mu_2 \sum_{n=0}^{\infty} b_n = \mu_1 \sum_{n=0}^{\infty} a_n$, we obtain

$$\sum_{n=0}^{\infty} b_n = \rho_2, \ \sum_{n=0}^{\infty} b'_n = q\rho_0, \sum_{n=0}^{\infty} a_n = \rho_1. \tag{5}$$

For state (M, n), $n \geq 1$, we have $m_n = (\frac{\rho_0}{1+\rho_0})^n m_0$. Therefore $\sum\limits_{i=0}^{\infty} m_i = m_0(1+\rho_0)$. Using now Equation (5) together with the normalization condition implies $m_0 = \frac{p}{1+p\rho_0}(1-\rho_1-\rho_2-\rho_3-q\rho_0)$. Equation (4) then becomes

$$P_0 = \frac{1-p}{1+p\rho_0}(1-\rho_1-\rho_2-q\rho_0-\rho_3)).$$

A new call enters service immediately upon arrival, if and only if the system is in state $(0,0)$. Since the call arrival process is Poison, we use the PASTA property to state that the steady-state probabilities seen by a new call arrival coincide with those seen at an arbitrary instant. Thus $P_D = 1 - P_0$, which leads to the expression of P_D.

As for the outbound job expected throughput, it is given by $T = \mu_0 \left(q \sum_{i=0}^{\infty} b_i + \sum_{i=0}^{\infty} b'_i + \sum_{i=0}^{\infty} m_i\right)$, which may be also written as $T = \mu_0 \left(\frac{1+\rho_0}{1+p\rho_0} p\,(1-\rho_1-\rho_2-\rho_3-q\rho_0) + q(\rho_2+\rho_0)\right)$. This finishes the proof of the proposition. □

Let us now define W, a random variable, as the steady-state call waiting time in the queue, and $\mathrm{P}(W \leq t)$ as its cumulative distribution function (cdf) for $t \geq 0$. Conditioning on a state seen by a new call arrival and averaging over all possibilities, we state using PASTA that

$$\begin{aligned}\mathrm{P}(W \leq t) = P_0 \cdot 1 + \sum_{n=0}^{+\infty} (&\mathrm{P}(W \leq t|(A,n)) \cdot a_n + \mathrm{P}(W \leq t|(B,n)) \cdot b_n + \mathrm{P}(W \leq t|(B',n)) \cdot b'_n) \\ &+ \mathrm{P}(W \leq t|(C,n)) \cdot c_n + \mathrm{P}(W \leq t|(M,n)) \cdot m_n).\end{aligned} \tag{6}$$

For $n \geq 0$, the quantities $\mathrm{P}(W \leq t|(A,n))$, $\mathrm{P}(W \leq t|(B,n))$, $\mathrm{P}(W \leq t|(B',n))$, $\mathrm{P}(W \leq t|(C,n))$ and $\mathrm{P}(W \leq t|(M,n))$ are the cdf of the conditional call waiting times in the queue, given that a new arriving call finds the system in states (A,n), (B,n), (B',n), (C,n) and (M,n), respectively. In the Markov chain of Model PM, these conditional random variables correspond to first passage times to state $(0,0)$ starting from the system state upon a new call arrival. They are convolutions of independent exponential random variables with arbitrarily rates, not necessarily all equal or all

distinct. Using the expressions of the cdf of an hypoexponential distribution (Amari and Misra (1997), Legros and Jouini (2015)), we can explicitly derive the expressions of $\mathrm{P}(W \leq t|(A,n))$, $\mathrm{P}(W \leq t|(B,n))$, $\mathrm{P}(W \leq t|(B',n))$, $\mathrm{P}(W \leq t|(C,n))$ and $\mathrm{P}(W \leq t|(M,n))$, for $n \geq 0$, as shown in Section 4 of the online appendix.

It remains now to compute the probabilities a_n, b_n, b'_n, c_n and m_n in n, for $n \geq 0$. One can compute them using the well-known matrix geometric solution approach (Neuts, 1995). However, the numerical computation is not exact, because the minimal non-negative solution to the matrix quadratic equation is computed with a given error. In what follows, we instead use the Cardan-Ferrari method to solve the moment generating function and find the roots, which leads to an exact numerical computation. From the Markov chain of Model PM, we can write the following iterative equations

$$\lambda X_{n-1} = G X_n, \tag{7}$$

for $n \geq 1$, where

$$X_n = \begin{pmatrix} a_n \\ b_n \\ b'_n \\ c_n \\ m_n \end{pmatrix},$$

for $n \geq 0$ is the vector of probabilities to be computed and

$$G = \begin{pmatrix} \mu_1 & -\lambda & -\lambda & -\lambda & -\lambda \\ -\mu_1 & \lambda+\mu_2 & 0 & 0 & 0 \\ 0 & -q\mu_2 & \lambda+\mu_0 & 0 & 0 \\ 0 & -(1-q)\mu_2 & -\mu_0 & \lambda+\mu_3 & 0 \\ 0 & 0 & 0 & 0 & \lambda+\mu_0 \end{pmatrix}.$$

The first step to solve Equation (7) is to find the eigenvalues of the matrix $\frac{1}{\lambda}G$. These are solutions of the equation $\det(\frac{1}{\lambda}G - yI) = 0$ with y as variable. One obvious eigenvalue is $1 + \frac{1}{\rho_0}$ (see the last line of G), and the remaining ones are those of a 4×4 matrix (derived from $\frac{1}{\lambda}G$ by removing the last line and the last column) and they are solutions of the following quartic equation

$$\sigma_4 y^4 - (3\sigma_4 + \sigma_3)y^3 + (3\sigma_4 + 2\sigma_3 + \sigma_2)y^2 - (\sigma_4 + \sigma_3 + \sigma_2 + \sigma_1)y + 1 + \rho_0(1-q) = 0, \tag{8}$$

with y as variable, $\sigma_1 = \rho_0 + \rho_1 + \rho_2 + \rho_3$, $\sigma_2 = \rho_0\rho_1 + \rho_0\rho_2 + \rho_0\rho_3 + \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3$, $\sigma_3 = \rho_0\rho_1\rho_2 + \rho_0\rho_1\rho_3 + \rho_0\rho_2\rho_3 + \rho_1\rho_2\rho_3$, and $\sigma_4 = \rho_0\rho_1\rho_2\rho_3$. Since the constant term $1 + \rho_0(1-q)$ in Equation (8) is strictly positive, zero cannot be a solution of that equation. Then, $\frac{1}{\lambda}G$ is invertible.

Therefore the eigenvalues of λG^{-1} are solutions of

$$(1+\rho_0(1-q))x^4 - (\sigma_4+\sigma_3+\sigma_2+\sigma_1)x^3 + (3\sigma_4+2\sigma_3+\sigma_2)x^2 - (3\sigma_4+\sigma_3)x + \sigma_4 = 0, \qquad (9)$$

where $x = \frac{1}{y}$. We solve the quartic Equation (9) using the Cardan-Ferrari method. In Section 5 of the online appendix, we describe the details of this method.

The explicit expressions of the probability components of the vector X_n, for $n \geq 0$, can be derived, however they are too cumbersome for Model PM. We go further in providing their expressions for the extreme cases of Model PM in Section 6 of the online appendix, and also using a light-traffic approximation in Section 7 of the online appendix. In all cases, an exact numerical method is straightforward and easy to implement. Numerical illustrations are shown later in Section 4.2.

Let us now compute the expected call waiting time in Model PM, denoted by $E(W)$. Consider first a model similar to Model PM except that outbound jobs can only be treated inside a call conversation. We denote this model by Model PM', and its call expected waiting time by $E(W')$. With a little thought, one can see that the expected call waiting time in Model PM is that of Model PM' plus $\frac{p}{\mu_0}$. The reason is mainly related to the memoryless property of outbound job service times. This result is proven in Lemma 1.

Lemma 1 *The expected waiting time in PM is delayed by $\frac{p}{\mu_0}$ compared to that in PM', for $p \in [0,1]$.*

Proof. See Section 8 of the online appendix. □

Note that the proof of Lemma 1 also holds for all independent and identically generally distributed call service times. Let us now compute the expected waiting time in Model PM', $E(W')$. We use the Pollaczeck-Kinchin result for an M/G/1 queue. From Pollaczeck (1930), we have $E(W') = \frac{\rho^2(1+c_v^2)}{2\lambda(1-\rho)}$, where c_v is the coefficient of variation of the service distribution (standard deviation over expected value) and ρ is the server utilization (expected arrival rate over expected service rate). Because of the possibility to do outbound jobs between calls, the random variable measuring the service time duration, say S, can be written as $S = S_1 + S_2 + US_0 + S_3$, where S_i, a random variable, follows an exponential distribution with rate μ_i, for $i = 0, ..., 3$, and U follows a Bernoulli distribution with parameter q. We denote by $E(Z)$ and $V(Z)$ the expected value (first moment) and the variance of a given random variable Z, respectively. The first moment of S is given by

$$E(S) = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{q}{\mu_0} + \frac{1}{\mu_3},$$

and its variance can be written as (using the independence between S_i and S_j for $i \neq j \in \{0, ..., 3\}$)

$$V(S) = V(S_1) + V(S_2) + V(US_0) + V(S_3) = \frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{2q - q^2}{\mu_0^2} + \frac{1}{\mu_3^2}.$$

Then

$$c_v^2(S) = \frac{\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{2q-q^2}{\mu_0^2} + \frac{1}{\mu_3^2}}{\left(\frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{q}{\mu_0} + \frac{1}{\mu_3}\right)^2}.$$

After some algebra, we obtain

$$E(W') = \frac{(\rho_1 + \rho_2 + \rho_3)^2 + \rho_1^2 + \rho_2^2 + \rho_3^2 + 2q\rho_0(\rho_0 + \rho_1 + \rho_2 + \rho_3)}{2\lambda(1 - (\rho_1 + \rho_2 + q\rho_0 + \rho_3))},$$

which leads to

$$E(W) = \frac{p}{\mu_0} + \frac{(\rho_1 + \rho_2 + \rho_3)^2 + \rho_1^2 + \rho_2^2 + \rho_3^2 + 2q\rho_0(\rho_0 + \rho_1 + \rho_2 + \rho_3)}{2\lambda(1 - (\rho_1 + \rho_2 + q\rho_0 + \rho_3))}. \tag{10}$$

This closes the performance measure analysis of Model PM. Note that for the waiting time of inbound calls, we ignored the waiting time they may have before starting stage 3. We expect that the waiting experience is more comfortable before stage 3 than before starting service. The customer sensitivity to uncertain waiting should decrease after being connected to an agent, because she would consider such waiting time as a part of her service time. For instance, one could expect that abandonments would be higher while waiting before stage 1 than before stage 3. For the rest of the paper, waiting before stage 3 is ignored. We however provide in Section 9 of the online appendix the details to characterize the distribution of the whole waiting time including that before stage 3.

For the four extreme cases of Model PM (Models 1,...,4), we derive the expressions of the outbound job expected throughput, the call probability of delay, and the call expected waiting time, we simply apply the previous analysis and state the results as shown in Table 2.

4.2 Comparison Analysis and Insights

We start in Section 4.2.1 by a comparison analysis between the extreme cases Models 1,...,4. The comparison is based on the optimization problem (1). We derive various structural results and properties for this comparison. In particular, we investigate the impact of the mean arrival rate intensity of calls on the comparison between Models 1,...,4. One could think of a call center manager that adjusts the job routing schema as a function of the call arriving workload over the day. In Section 4.2.2 we focus on the general case Model PM. We prove that the optimization

Table 2: Expressions of T, $E(W)$ and P_D for Models 1,...,4

	Model 1	Model 2
T	0	$\mu_0(1-\rho_1-\rho_2-\rho_3)$
$E(W)$	$\frac{(\rho_1+\rho_2+\rho_3)^2(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2}{(\rho_1+\rho_2+\rho_3)^2})}{2\lambda(1-\rho_1-\rho_2-\rho_3)}$	$\frac{1}{\mu_0}+E(W_1)$
P_D	$\rho_1+\rho_2+\rho_3$	1
	Model 3	**Model 4**
T	$\mu_0(\rho_0+\rho_2)$	$\mu_0(1-\rho_1-\rho_3)$
$E(W)$	$\frac{(\rho_1+\rho_2+\rho_3+\rho_0)^2(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2+\rho_0^2}{(\rho_1+\rho_2+\rho_3+\rho_0)^2})}{2\lambda(1-\rho_1-\rho_2-\rho_3-\rho_0)}$	$\frac{1}{\mu_0}+E(W_3)$
P_D	$\rho_0+\rho_1+\rho_2+\rho_3$	1

of the parameters of Model PM lead to extreme situations in the sense of a systematic outbound job treatment of outbound jobs either between calls or inside a call conversation, which gives an interest in practice for Models 1,...,4.

4.2.1 Comparison between the Extreme Cases

We first compare between Models 1,...,4 based on their performance in terms of the outbound job expected throughput, denoted by T_1,..., T_4, respectively. It is obvious that Model 4 is the best and Model 1 is the worst (no outbound jobs at all). Let us now compare between Models 2 and 3. From Table 2 we have $T_2 = \mu_0(1-\rho_1-\rho_2-\rho_3)$ and $T_3 = \mu_0(\rho_0+\rho_2)$. Thus $T_3 > T_2$ is equivalent to

$$\lambda > \frac{1}{\frac{1}{\mu}+\frac{1}{\mu_2}},$$

where $\frac{1}{\mu} = \frac{1}{\mu_0}+\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}$. Since the stability condition for Model 3 is $\lambda < \mu$, Model 3 is better than Model 2 if

$$\frac{1}{\frac{1}{\mu}+\frac{1}{\mu_2}} < \lambda < \mu. \tag{11}$$

Denoting the left term in Inequality (11) by R, the condition under which $T_3 > T_2$ is then

$$R = \frac{1}{\frac{1}{\mu_0}+\frac{1}{\mu_1}+\frac{1}{\mu_3}+\frac{2}{\mu_2}} < \lambda < \mu. \tag{12}$$

From Inequality (12), we first see that treating outbound jobs only inside a call conversation (Model 3) becomes better than treating them only between calls (Model 2) is likely the case for high arrival workloads (in such a case, idle period durations are reduced). We also see that $\frac{\partial R}{\partial \mu_2} > 0$ for $\mu_2 > 0$, $\frac{\partial R}{\partial \mu_0} > 0$ for $\mu_0 > 0$, $\frac{\partial R}{\partial \mu_1} > 0$ for $\mu_1 > 0$, and $\frac{\partial R}{\partial \mu_3} > 0$ for $\mu_3 > 0$. This means that

decreasing the expected duration of the call service second stage ($1/\mu_2$) relative to the expected durations of the other call service stages or the outbound job service duration ($1/\mu_1$, $1/\mu_3$ and $1/\mu_0$) increases the range of arrival workloads where it is preferred to use Model 2 instead of Model 3. In other words, there is no sufficient time to treat outbound jobs inside the call conversation.

Comparison with a Constraint on $E(W)$

As a function of the mean call arrival rate, we want to answer the question when should we treat outbound jobs (which model among Models 1 to 4 should a manger choose?) for the following problem

$$\begin{cases} \text{Maximize } T \\ \text{subject to } E(W) \leq w^*, \end{cases} \tag{13}$$

where w^* is the service level for the expected waiting time, $w^* > 0$. Let W_i, a random variable, denote the expected call waiting time in Model i, $i = 1, ..., 4$. It is clear that for some periods of a working day with a very high call arrival rate λ, the manager is likely to choose Model 1 (no outbound jobs), and for other periods with a very low λ, she is likely to choose Model 4 (outbound jobs between calls and inside a call). However for intermediate values of λ, the optimal choice is not clear. This is what we analytically analyze in what follows.

Under the condition of stability of Model i, $E(W_i)$ is continuous and strictly increasing in λ (see Table 1), for $i = 1, ..., 4$. The constraint $E(W_i) \leq w^*$ is then equivalent to $\lambda \leq \overline{\lambda}_i$, for $i = 1, ..., 4$, where

$$\begin{aligned}
\overline{\lambda}_1 &= \frac{2w^*}{2w^*\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=1}^{3}\frac{1}{\mu_i^2}}, \\
\overline{\lambda}_2 &= \frac{2\left(w^* - \frac{1}{\mu_0}\right)}{2\left(w^* - \frac{1}{\mu_0}\right)\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=1}^{3}\frac{1}{\mu_i^2}}, \\
\overline{\lambda}_3 &= \frac{2w^*}{2w^*\left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=0}^{3}\frac{1}{\mu_i^2}}, \\
\overline{\lambda}_4 &= \frac{2\left(w^* - \frac{1}{\mu_0}\right)}{2\left(w^* - \frac{1}{\mu_0}\right)\left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=0}^{3}\frac{1}{\mu_i^2}}.
\end{aligned} \tag{14}$$

For a given λ and under the condition of stability of Model i ($i = 1, ..., 4$), the choice of Model i happens if $\lambda \leq \overline{\lambda}_i$ and $T_i = \max_{j \in \{1,...,4\},\ \lambda \leq \overline{\lambda}_j}(T_j)$. When $\lambda \leq \overline{\lambda}_4$, the choice is obviously for Model 4. When $\lambda \leq \overline{\lambda}_1$ and $\lambda > \overline{\lambda}_i$ for $i = 2, 3, 4$ the only possibility is Model 1. Proposition 2 provides

the conditions under which an optimal choice of Model 2 or Model 3 may happen.

Proposition 2 *The following holds:*

1. *For* $\lambda < \frac{1}{\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}}$, *there exist some values of* λ *for which Model 2 is optimal if and only if* $\overline{\lambda}_2 > 0$.

2. *For* $\lambda < \frac{1}{\frac{1}{\mu_0}+\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}}$, *there exist some values of* λ *for which Model 3 is optimal if and only if*

$$\begin{cases} R \leq \overline{\lambda}_3 \\ \ or \\ \overline{\lambda}_2 < \overline{\lambda}_3. \end{cases}$$

3. *We have* $\overline{\lambda}_2 < \overline{\lambda}_3$ *if and only if* $\frac{1}{\mu_0} < w^* < \overline{w}^*$, *where*

$$\overline{w}^* = \frac{1}{2}\sqrt{\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3}\frac{1}{\mu_i^2} + 6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j}} - \left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right).$$

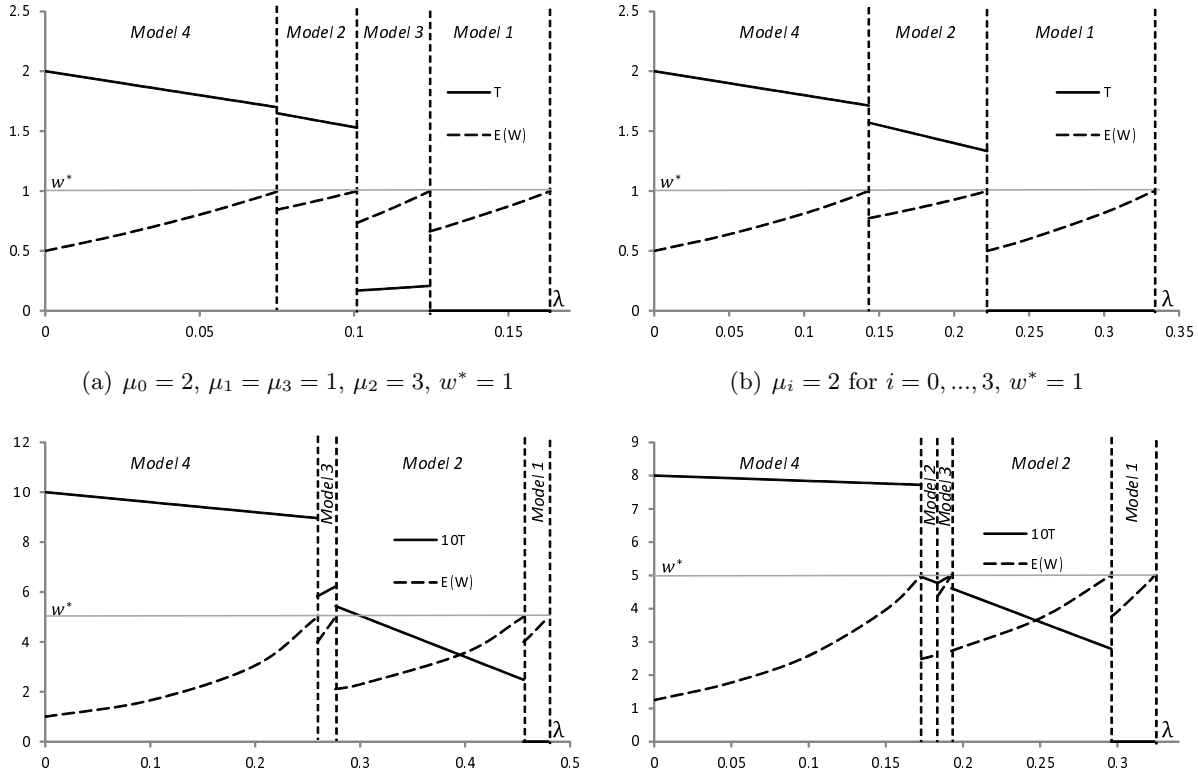
Proof. See Section 10 of the online appendix. □

Using Equation (14), the condition in the first statement of Proposition 2 may be rewritten as

$$\begin{cases} w^* > \frac{1}{\mu_0} \\ \text{ or } \\ w^* < \frac{1}{\mu_0} - \frac{\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=1}^{3}\frac{1}{\mu_i^2}}{2\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)}. \end{cases} \tag{15}$$

The second inequality in Relation (15) implies $w^* < \frac{1}{\mu_0}$. Since at least the expected waiting time in Model 2 is strictly higher than $\frac{1}{\mu_0}$ (any new call has at least to wait for the residual time of an outbound job treatment), this second inequality is impossible. Roughly speaking, the condition for the optimality of Model 2 (for some values of λ) holds when the service level on the call waiting is higher than the expected outbound job service time.

In what follows, we numerically illustrate the analysis above. For various system parameters, Figure 5 gives the optimal model choice as a function of the mean arrival rate of calls, λ. An intuitive reasoning of a manager would choose the ordering Model 4 (outbound jobs between calls and inside a call), then 2 (outbound jobs only between calls), then 3 (outbound jobs only inside a call), then 1 (no outbound jobs) as λ increases.



(a) $\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $w^* = 1$

(b) $\mu_i = 2$ for $i = 0, ..., 3$, $w^* = 1$

(c) $\mu_0 = 1$, $\mu_1 = \mu_3 = 5$, $\mu_2 = 0.8$, $w^* = 5$

(d) $\mu_0 = 0.8$, $\mu_1 = \mu_3 = 10$, $\mu_2 = 0.5$, $w^* = 5$

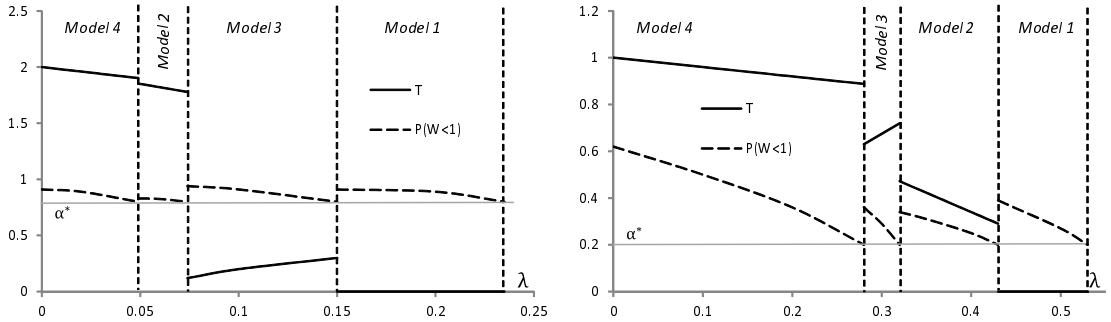
Figure 5: Comparison between Models 1 to 4 with a constraint on $E(W)$

The ordering Model 2 then Model 3 is not always appropriate, and some situations may require to consider some counterintuitive ordering. For instance, Model 3 is better than Model 2 for small values of λ if $R \leq \overline{\lambda}_4$ and $\overline{\lambda}_3 < \overline{\lambda}_2$, see Figure 5(c). In other words, this happens when the constraint on $E(W)$ is not too restrictive and when the expected second stage service duration is long. Another more surprising ordering, as λ increases, is Model 2, then Model 3, then again Model 2 (see Figure 5(d)) which happens for system parameters such that $\overline{\lambda}_4 < R < \overline{\lambda}_3 < \overline{\lambda}_2$.

Comparison with a Constraint on $P(W < A_{WT})$

In the constraint of Problem (1), we want that the probability that a call waits less than a given threshold, defined as A_{WT} is at least a given service level, defined as SL, i.e., $\mathrm{P}(W < A_{WT}) \geq SL$. Note that a special case of this constraint is that on P_D, the call probability of waiting. The expressions involved in the analysis of $P(W < A_{WT})$ are quite complicated to allow an analytical comparison between the models as we have done for a constraint on $E(W)$. We have then conducted a numerical comparison analysis (not totally illustrated here). The main qualitative conclusions are similar to those for the case of a constraint on $E(W)$. As λ increases, it is not always true as one would intuitively expect that a manager should choose first Model 2 and then at some point of λ she shifts to Model 3 (Figure 6(a)). The optimal choice may change with the system parameters

and we may have the ordering Model 3 then Model 2 (Figure 6(b)).



(a) $\mu_i = 2$ for $i = 0, ..., 3$ and $\alpha^* = 80\%$ (b) $\mu_0 = 1$, $\mu_1 = 5 = \mu_3$, $\mu_2 = 0.8$ and $\alpha^* = 20\%$

Figure 6: Comparison between Models 1 to 4 with the constraint on $\mathrm{P}(W < 1)) \geq \alpha^*$

4.2.2 Optimization of Model PM

In this section we focus on the general case, Model PM. We are interested in the optimization of the parameters p and q in Model PM for Problem (1). Concretely, we want to find the optimal routing parameters of Model PM that allows the manager to maximize the outbound job expected throughput while respecting a call service level constraint. Recall that the stability condition of Model PM is $\lambda < \frac{1}{\frac{q}{\mu_0}+\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}}$.

The expression of the outbound job expected throughput T for Model PM is given in Proposition 1. It is straightforward to prove that for $p, q \in [0, 1]$ the maximum of T (best situation) is reached for $p = q = 1$. The proof is then omitted. Also, the expected call waiting time of Model PM (given in Equation (10)) is maximized (worst) for $p = q = 1$. Therefore in order to solve Problem (1), one would be interested analyzing the sensitivity of T with respect to p and q. In Lemma 2 we prove a sensitivity result for T. The result will be used later in our analysis.

Lemma 2 *We have* $\frac{\partial T}{\partial p} > \frac{\partial T}{\partial q}$ *if and only if* $0 < \rho_0 < \overline{\rho}_0$, *where*

$$\overline{\rho}_0 = \frac{\sqrt{(p-q-\rho_1-\rho_2-\rho_3)^2 - 4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3)} - q + p - (\rho_1+\rho_2+\rho_3)}{2(q-p^2+p)}.$$

Proof. We want to solve the following inequality in ρ_0:

$$\frac{\partial T}{\partial p} = \mu_0 \frac{(1-\rho_1-\rho_2-q\rho_0-\rho_3)(1+\rho_0)}{(1+p\rho_0)^2} > \frac{\partial T}{\partial q} = \mu_0 \frac{\rho_0(1-p)}{1+p\rho_0}.$$

This is equivalent to $(1-\rho_1-\rho_2-\rho_3-q\rho_0)(1+\rho_0) - \rho_0(1-p)(1+p\rho_0) > 0$, or also

$$(p^2-p-q)\rho_0^2 + (p-q-\rho_1-\rho_2-\rho_3)\rho_0 + 1 - \rho_1 - \rho_2 - \rho_3 > 0. \tag{16}$$

The discriminant for this inequality is $\Delta = (p-q-\rho_1-\rho_2-\rho_3)^2 - 4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3) > 0$.

Equation (16) has then the two following solutions:

$$\frac{\sqrt{(p-q-\rho_1-\rho_2-\rho_3)^2-4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3)}-q+p-(\rho_1+\rho_2+\rho_3)}{2(q-p^2+p)}$$

and

$$-\frac{\sqrt{(p-q-\rho_1-\rho_2-\rho_3)^2-4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3)}+q-p+\rho_1+\rho_2+\rho_3}{2(q-p^2+p)}.$$

Since the first solution is positive (denoted by $\overline{\rho}_0$) and the second one is negative, $\rho_0 \in [0;\rho_0^*)$, which finishes the proof of the lemma. □

In what follows we address the question: starting from $p=q=1$, in which direction should we decrease T? Should we decrease p or q first?
For $p=q=1$, we have $\overline{\rho}_0 = \frac{1}{2}\left(\sqrt{(\rho_1+\rho_2+\rho_3)^2+4(1-\rho_1-\rho_2-\rho_3)}-(\rho_1+\rho_2+\rho_3)\right)$. Let us now prove (for $p=q=1$) that $\overline{\rho}_0 > \rho_0$. From the one hand, proving $\overline{\rho}_0 > \rho_0$ is equivalent to proving $\sqrt{(\rho_1+\rho_2+\rho_3)^2+4(1-\rho_1-\rho_2-\rho_3)} > 2\rho_0+(\rho_1+\rho_2+\rho_3)$ or equivalently $\rho_0^2+\rho_0(\rho_1+\rho_2+\rho_3)-(1-(\rho_1+\rho_2+\rho_3))<0$ or also $(\rho_0+1)(\rho_0-(1-(\rho_1+\rho_2+\rho_3)))<0$. From the other hand, we have $\rho_0+1>0$. Also, the stability condition of Model 4 (Model PM with $p=q=1$) is $\rho_0+\rho_1+\rho_2+\rho_3<1$. Then $\rho_0<1-(\rho_1+\rho_2+\rho_3)$. As a conclusion the inequality $\overline{\rho}_0>\rho_0$ is true, for $p=q=1$. Using Lemma 2, this means that starting from $p=q=1$, we have $\frac{\partial T}{\partial p} > \frac{\partial T}{\partial q} > 0$. As a consequence, when we need to modify the values of p and q in order to decrease the expected call waiting time (the constraint in Problem (1)), the maximum of T is guaranteed by first decreasing q (the outbound job expected throughput is less sensitive to the variation of q than that of p). The question now is: which direction to use next? In other words when $p=1$ and some value of q such that $0<q<1$, is it possible that it is better to decrease p instead of q? The answer is no and the proof is as follows. For $p=1$, let us try to find a value of q for which we have $\overline{\rho}_0 \leq \rho_0$. This is equivalent to $\frac{\sqrt{(1-q-\rho_1-\rho_2-\rho_3)^2+4q(1-\rho_1-\rho_2-\rho_3)}-q+1-\rho_1-\rho_2-\rho_3}{2q} \leq \rho_0$. Thus, $q^2\rho_0^2+q\rho_0-(1-\rho_1-\rho_2-\rho_3)(1+\rho_0)>0$. This trinomial in q has two real solutions; $q_1 = -\frac{1+\sqrt{4\rho_0+5-4(\rho_1+\rho_2+\rho_3)(\rho_0+1)}}{2\rho_0}$ and $q_2 = \frac{-1+\sqrt{4\rho_0+5-4(\rho_1+\rho_2+\rho_3)(\rho_0+1)}}{2\rho_0}$. Obviously $q_1<0$. We also have $q_2>1$ because: proving $q_2-1>0$ is equivalent to proving $\rho_0^2+(\rho_1+\rho_2+\rho_3)\rho_0+1>0$. The discriminant of this latter trinomial in ρ_0 is negative and it is equal to $(\rho_1+\rho_2+\rho_3)^2-4$. So $q_2>1$ for any $\rho_0>0$. Therefore it is impossible to find a value of q between 0 and 1 for which $0<\frac{\partial T}{\partial p}<\frac{\partial T}{\partial q}$. In conclusion starting from $p=q=1$, when we need to change the values of p and q, the best direction to maximize T is to first decrease q until $q=0$ and only then start to decrease p from $p=1$.

Consider now Problem (1) with a constraint on $E(W)$. From the one hand, the constraint

$E(W) \leq w^*$ implies

$$\frac{p}{\mu_0} + \frac{(\rho_1+\rho_2+\rho_3)^2+\rho_1^2+\rho_2^2+\rho_3^2+2q\rho_0(\rho_0+\rho_1+\rho_2+\rho_3)}{2\lambda(1-(\rho_1+\rho_2+q\rho_0+\rho_3))} \leq w^*,$$

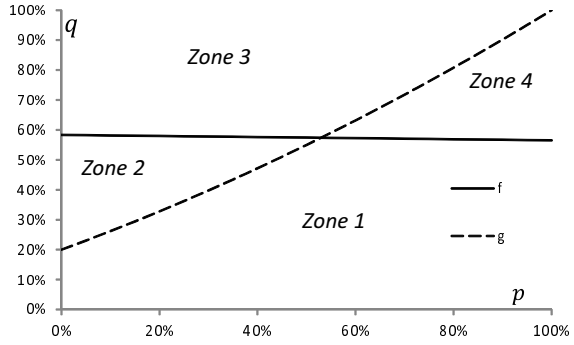
for $p, q \in [0,1]$, or equivalently

$$q \leq \frac{2\lambda(1-\rho_1-\rho_2-\rho_3)(w^*-p/\mu_0)-(\rho_1+\rho_2+\rho_3)^2-\rho_1^2-\rho_2^2-\rho_3^2}{2\rho_0(\rho_0+\rho_1+\rho_2+\rho_3+\lambda(w^*-p/\mu_0))}, \tag{17}$$

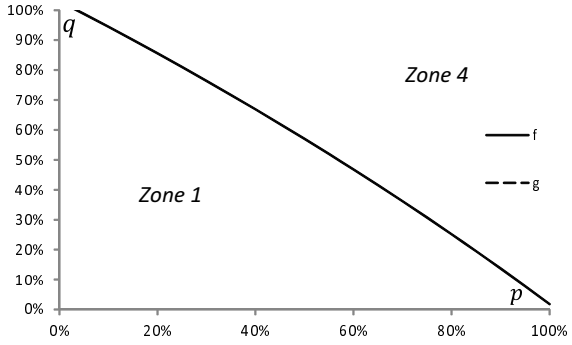
for $p, q \in [0,1]$. On the other hand, the condition in Lemma 2, $0 < \rho_0 < \overline{\rho}_0$, is equivalent to

$$q < \frac{1-(\rho_1+\rho_2+\rho_3)(1+\rho_0)}{\rho_0(1+\rho_0)} + \frac{1-\rho_0}{1+\rho_0}p + \frac{\rho_0}{1+\rho_0}p^2, \tag{18}$$

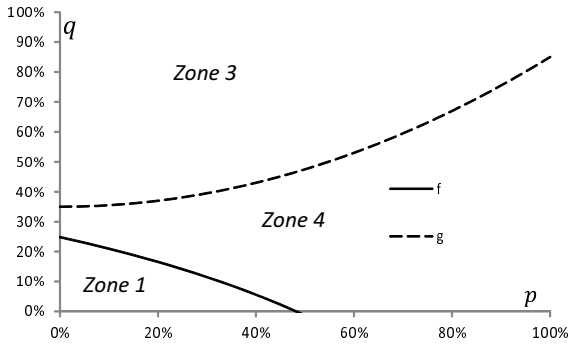
for $p, q \in [0,1]$. Let us denote the right hand sides of Inequalities (17) and (18) by the functions in $p \in [0,1]$ $f(p)$ and $g(p)$, respectively. Illustrations of these functions are given in Figure 7.



(a) $\lambda = 0.5$, $\mu_0 = \mu_1 = \mu_2 = \mu_3 = 2$, $w^* = 10$



(b) $\lambda = 0.22$, $\mu_0 = \mu_1 = \mu_2 = \mu_3 = 2$, $w^* = 1$



(c) $\lambda = 0.1$, $\mu_0 = 0.1$, $\mu_1 = \mu_2 = \mu_3 = 2$, $w^* = 5$

Figure 7: Behavior of $f(p)$ and $g(p)$

In what follows we prove an interesting result on the optimal values of p and q. Consider for example Figure 7(a) and assume that the agent is in a situation such that (p, q) belongs to Zone 1 or 2. Then the constraint on $E(W)$ is respected, but T can be improved. Using Lemma 2, we should increase p first (q first) for Zone 1 (Zone 2). From Figure 7(a), we also see that we should

decrease p first (q first) for Zone 3 (Zone 4). It is clear that the optimal couple (p, q) will be on the curve of f. Moreover, we prove in Theorem 1 that the optimal point is such that $p \in \{0, 1\}$ or $q \in \{0, 1\}$.

Theorem 1 *For $p, q \in [0, 1]$, the optimal values of p and q of the optimization problem*

$$\begin{cases} \textit{Maximize } T \\ \textit{subject to } E(W) \leq w^*, \end{cases} \tag{19}$$

are always extreme values (0 or 1) for at least p or q.

Proof. We want to maximize the outbound job expected throughput $T(p, q)$ while respecting a constraint on the expected call waiting time ($E(W)(p, q) \leq w^*$). We distinguish the following cases.
Case 1: $\lambda w^* < \frac{(\rho_1+\rho_2+\rho_3)^2\left(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2}{(\rho_1+\rho_2+\rho_3)^2}\right)}{2(1-\rho_1-\rho_2-\rho_3)}$**.** In this case the constraint on the expected waiting time cannot be met and Problem (1) has no solution.
Case 2: $\lambda w^* \geq \rho_0 + \frac{(\rho_1+\rho_2+\rho_3+\rho_0)^2\left(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2+\rho_0^2}{(\rho_1+\rho_2+\rho_3+\rho_0)^2}\right)}{2(1-\rho_1-\rho_2-\rho_3-\rho_0)}$**.** Since $T(p, q)$ increasing in both p and in q and the constraint on $E(W)$ is met for $p = q = 1$, the optimal values for the control parameters are $p = q = 1$ (Model 4).
Case 3: $\frac{(\rho_1+\rho_2+\rho_3)^2\left(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2}{(\rho_1+\rho_2+\rho_3)^2}\right)}{2(1-\rho_1-\rho_2-\rho_3)} < \lambda w^* < \rho_0 + \frac{(\rho_1+\rho_2+\rho_3+\rho_0)^2\left(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2+\rho_0^2}{(\rho_1+\rho_2+\rho_3+\rho_0)^2}\right)}{2(1-\rho_1-\rho_2-\rho_3-\rho_0)}$**.** In this case, since T is increasing in p and in q, the constraint on $E(W)$ should be saturated. We use the method of Lagrange multipliers to find the optimal point (p, q). Let us denote the Lagrange multiplier by α (α is real). Then α and the extremum (p, q) of our optimization problem are solutions of the set of the 3 equations $D(T + \alpha(W - w^*)) = 0$, where D is the differential applicator in α, p and q. These 3 equations are

$$\frac{\partial(T + \alpha(W - w^*))}{\partial p} = \mu_0 \frac{(1 - \rho_1 - \rho_2 - q\rho_0 - \rho_3)(1 + \rho_0)}{(1 + p\rho_0)^2} + \alpha\frac{1}{\mu_0} = 0, \tag{20}$$

$$\frac{\partial(T + \alpha(W - w^*))}{\partial q} = \mu_0 \frac{(1 - p)\rho_0}{1 + p\rho_0} + \alpha\frac{1}{2\lambda}\frac{\rho_0(2(\rho_0 + \rho_1 + \rho_2 + \rho_3)(1 - (\rho_1 + \rho_2 + \rho_3)) + (\rho_1 + \rho_2 + \rho_3)^2 + \rho_1^2 + \rho_2^2 + \rho_3^2)}{(1 - (\rho_1 + \rho_2 + \rho_3 + q\rho_0))^2} = 0, \tag{21}$$

$$\frac{\partial(T + \alpha(W - w^*))}{\partial \alpha} = \frac{p}{\mu_0} + \frac{(\rho_1 + \rho_2 + \rho_3)^2 + \rho_1^2 + \rho_2^2 + \rho_3^2 + 2q\rho_0(\rho_0 + \rho_1 + \rho_2 + \rho_3)}{2\lambda(1 - (\rho_1 + \rho_2 + q\rho_0 + \rho_3))} - w^* = 0. \tag{22}$$

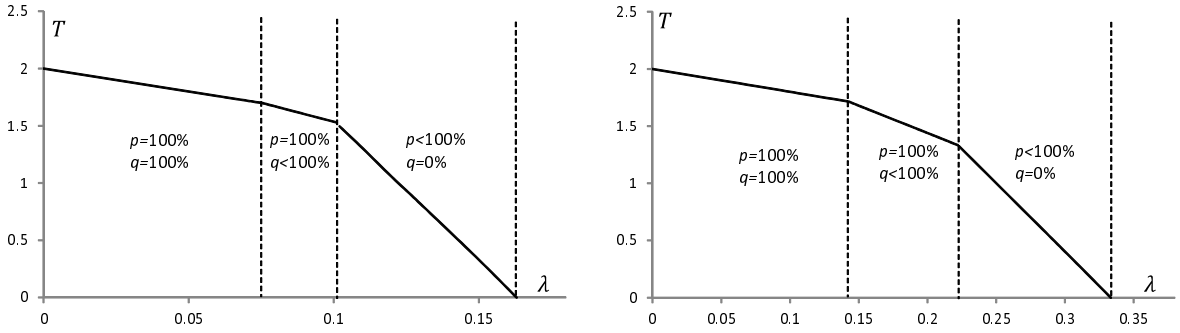
From Equation (20), we obtain $\alpha = -\mu_0^2 \frac{(1+\rho_0)(1-(\rho_1+\rho_2+\rho_3+q\rho_0))}{(1+p\rho_0)^2}$. Substituting this expression in Equation (21) leads to

$$\frac{(1 - p)\rho_0}{1 + p\rho_0} - \frac{(1 + \rho_0)(2(\rho_1 + \rho_2 + \rho_3 + \rho_0)(1 - (\rho_1 + \rho_2 + \rho_3)) + (\rho_1 + \rho_2 + \rho_3)^2 + \rho_1^2 + \rho_2^2 + \rho_3^2)}{2(1 + p\rho_0)^2(1 - (\rho_1 + \rho_2 + \rho_3 + q\rho_0))} = 0,$$

which implies

$$q = -\frac{(1-(\rho_1+\rho_2+\rho_3))(\rho_0^2(1+p^2)+\rho_0 p)+(\rho_1+\rho_2+\rho_3-\rho_1\rho_2-\rho_1\rho_3-\rho_2\rho_3)(1+\rho_0)}{\rho_0^2(1-p)(1+p\rho_0)}.$$

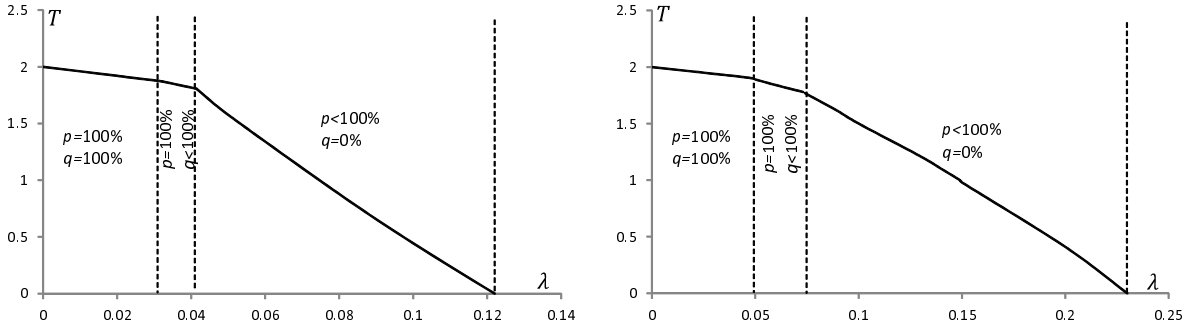
Since $1-(\rho_1+\rho_2+\rho_3)>0$ (stability condition for Model 1), we have $q<0$ if $p\in(0,1)$. It is therefore impossible to have a critical point with both p and q in $(0,1)$. So, the optimal solution of Problem (1) is not a critical point. We then deduce for the optimal values of p and q that at least one them needs to have an extreme value (0 or 1). This finishes the proof of the theorem. □



(a) $\mu_0=2$, $\mu_1=1=\mu_3$, $\mu_2=3$ (b) $\mu_i=2$ for $i=0,...,3$

Figure 8: Optimal p and q with $w^*=1$

Figure 8 provides a numerical illustration of Theorem 1. We also observe that as the arrival rate increases, it is optimal to first reduce the use of the break and next reduce the use of the time between the service of inbound calls. From Figure 9, we observe that the result in Theorem 1 with a constraint on the expected waiting time still holds with a constraint on the waiting time percentile (a rigorous proof is too complex).



(a) $\mu_0=2$, $\mu_1=\mu_3=1$, $\mu_2=3$ (b) $\mu_i=2$ for $i=0,...,3$

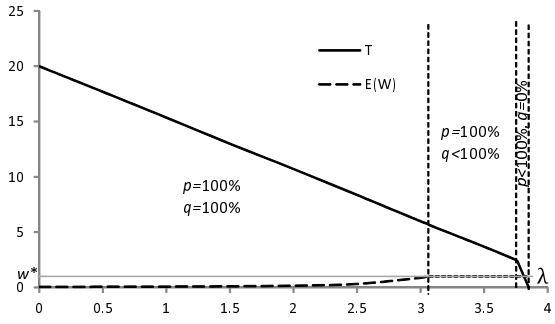
Figure 9: Optimal p and q with $\mathrm{P}(W<1)\geq 80\%$

5 Multi-Server Case

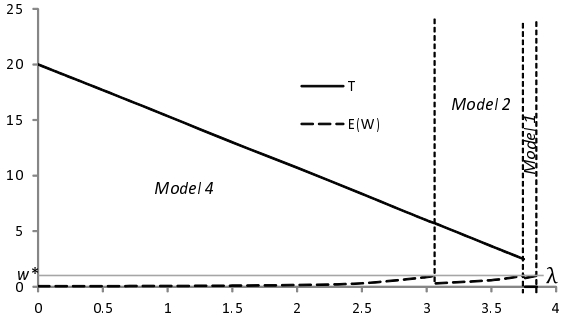
In this section, we focus on Problem (1) for the multi-server case. The idea is to assess the applicability of our single server results to a multi-server setting. We want to either optimize p and q for Model PM, or give the ordering of the extreme cases Models 1 to 4. An exact analysis as that done for the single server case is too complex. First, we conduct a simulation study to optimize (p, q) through an exhaustive search and relate the observations with the results in the single server case. We also investigate the impact of the system size on the optimal choice of p and q. Finally, we propose closed-form expressions for the approximate performance measures under light and heavy-traffic regimes. This allows to easily optimize the parameters p and q under those particular regimes.

5.1 Impact of the System Size

We use simulation to obtain the (p, q) couple which answers Problem (1). The optimization approach consists of an exhaustive search by discretizing the supports of p and q. The quality of the obtained solution is controlled through the choice of the spread between two adjacent discretized values. The results are given in Figure 10 and Table 3. Figure 10(a) provides a numerical evidence that Theorem 1 is still true for $s > 1$. We observe as a function of the system parameters that at least one of the routing parameters is either 0 or 1. This gives the merit to the study of the extreme cases Models 1, ..., 4. While increasing the workload, we again observe that the choice is first for $p = q = 1$; then $p = 1$ and $0 < q < 1$; then $0 < p < 1$ and $q = 0$. Thus the two questions of routing outbound jobs (between calls or during the break) are not considered together at the same time. Figure 10(b) also illustrates the ordering between the extreme cases.



(a) Optimal p and q



(b) Comparison between Models 1 to 4

Figure 10: Model choice with a constraint on $E(W)$ ($\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $w^* = 1$, $s = 10$)

Table 3 provides simulation results relating the number of agents s and the intervals of call arrival rates for the situations $p = q = 1$; $p = 1$ and $0 < q < 1$; $0 < p < 1$ and $q = 0$. The table gives the frontiers of the λ values that determine the choice of p and q. For example, the

Table 3: Impact of the system size ($\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $\mu_0 = 2$)

	λ		
s	$p = q = 1$	$p = 1, 0 < q < 1$	$0 < p < 1, q = 0$
1	0.08	0.10	0.16
5	1.34	1.66	1.74
10	3.06	3.75	3.85
20	6.56	8.04	8.09
50	17.08	20.83	20.87
100	35.08	42.50	42.53

line $s = 5$ indicates that the choice is $p = q = 1$ for $0 \leq \lambda \leq 1.34$; it is $p = 1$ and $0 < q < 1$ for $1.34 < \lambda \leq 1.66$; and it is $0 < p < 1$ and $q = 0$ for $1.66 < \lambda \leq 1.74$. Table 3 reveals that the interval of workload values for which the solution $p = q = 1$ answers Problem (1) enlarges in the system size. The explanation is related to the pooling effect. The larger is the system, the better are the performance for inbound calls. We then may profit from the two opportunities for the routing of outbound jobs (inside and between calls).

We also observe that the interval of workload values for which $p = 1$ enlarges in the system size. An explanation of this observation can be given using an approximation of the multi-server system. We propose a *super* server approximation as follows. We replace the s servers by a single server, where the service rates become $s\mu_0$, $s\mu_1$, $s\mu_2$, and $s\mu_3$ for outbound jobs, stage 1, stage 2 and stage 3, respectively. Although this approximation only works for the particular cases of very small number of servers or heavily loaded call centers, it still allows to provide some evidence of the numerical observation in the table. For the super server model, we can apply the single server results. Let us then consider the threshold call arrival rate values $\overline{\lambda}_i$, for $i = 1, ..., 4$, in Equations (14), and let us substitute the service rates μ_i by $s\mu_i$, for $i = 0, ..., 3$. These thresholds determine the limits of the intervals where a system manager should choose p and/or q equal to 0 or 1. For a large value of s, we have $\frac{1}{s\mu_0} << w^*$, $\frac{1}{s}\left(\sum_{i=j}^{3}\frac{1}{\mu_i}\right)^2 << \sum_{i=j}^{3}\frac{1}{\mu_i}$ and $\frac{1}{s}\sum_{i=j}^{3}\frac{1}{\mu_i^2} << \sum_{i=j}^{3}\frac{1}{\mu_i}$ for $j \in \{0, 1\}$. This leads to $\overline{\lambda}_1 = \overline{\lambda}_2 = \frac{s}{\sum_{i=1}^{3}\frac{1}{\mu_i}}$ and $\overline{\lambda}_3 = \overline{\lambda}_4 = \frac{s}{\sum_{i=0}^{3}\frac{1}{\mu_i}}$, which proves that the interval of λ for which Models 1 and 3 are considered can be neglected when s is large. Recall that the stability constraint for Models 1 and 2 is $\lambda < \frac{s}{\sum_{i=1}^{3}\frac{1}{\mu_i}}$. Thus, when s is high, Model 2 ($q = 0$ and $p = 1$) can be used until we hit the system instability. This means that in large call centers we should most of the time route outbound jobs between calls. An intuitive explanation is related to the high number of agents in large call centers. In such a case, even with a choice of $p = 1$ (i.e., systematic decision of initiating outbound jobs between inbound calls), an arriving call will not likely wait much longer than with a choice of $p < 1$.

The remaining question is the routing of outbound jobs during the break. To the contrary to

the routing between calls, there is not an asymptotic choice for the routing inside a call as the system size grows. Since $\frac{\overline{\lambda}_1-\overline{\lambda}_4}{\overline{\lambda}_4} = \frac{\frac{1}{\mu_0}}{\sum_{i=0}^{3}\frac{1}{\mu_i}}$, the relative length of the interval $(\overline{\lambda}_4, \overline{\lambda}_1)$ in which $p = 1$ and $0 \leq q \leq 1$ does not depend on s. This agrees with the numerical experiments in Table 3.

5.2 Approximations

We develop here an approximate analysis of the multi-server performance measures under light and heavy-traffic regimes of inbound calls. We also illustrate how the approximations can be used to solve Problem (1).

Light-traffic approximation. It is unusual to observe a light-traffic regime in traditional call centers with only inbound calls. The heavy-traffic regime is instead observed since the common practice is to make agents busy almost all the time. The main motivation for considering the light-traffic regime is that it can be decided by the management in a multi-channel context. The decision to overstaff for inbounds in order to treat a high volume of outbounds can be motivated by the value that outbounds may represent for the call center (case of sales activities for example). Note however that since the number outbound jobs is sufficiently large, even with a very light-traffic of inbound calls, agents may remain busy most of the time.

Proposition 3 provides the approximate performance measures under a light-traffic regime. We write $f(x) \underset{x\to x_0}{\sim} g(x)$ to state that $\lim_{x \to x_0} \frac{f(x)}{g(x)} = 1$, for $x_0 \in \mathbb{R}$.

Proposition 3 *We have* $E(W) \underset{\lambda\to 0}{\sim} \frac{p}{s\mu_0}$, $T \underset{\lambda\to 0}{\sim} \mathbb{I}_{p>0}(s-1)\mu_0 + p\mu_0$, *and for* $A_{WT} > 0$, $P(W < A_{WT}) \underset{\lambda\to 0}{\sim} 1 - pe^{-\mu_0 s A_{WT}}$, *where* $\mathbb{I}_{p>0}$ *is 1 for* $p > 0$ *and 0 otherwise.*

Proof. Assume that under a light-traffic regime, we never observe more than one inbound call in the system. Recall that an agent stops treating outbound jobs after the service completion of an outbound job only if all the other agents are busy and one call is waiting in the queue. Consider an empty system with idle agents. Upon the service completion of the first call, the involved agent, referred to as agent 1, starts the treatment of outbound jobs with probability p. Thus, upon the service completion of the first call, one agent is working on an outbound job and $s-1$ agents are idle with with probability p, and all agents are idle with probability $1-p$. Upon the arrival of the second call (the first call has already left the system, and agent 1 is still working on outbound jobs), this call is immediately routed to an idle agent. Upon the call service completion, again this agent starts the treatment of outbound jobs with a probability p. Thus, upon the service completion of the second call, two agents are working on outbound jobs and $s-2$ agents are idle with probability p^2, one agent is working on an outbound job and $s-1$ agents are idle with probability $2p(1-p)$, and all agents are idle with probability $(1-p)^2$. Thus, for $p > 0$, as the total number of arrived calls increases the probability that $s-1$ agents are working on outbound jobs converges to one.

Table 4: Light-traffic approximation ($\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $s = 10$, $q = 50\%$)

		Simulation			Approximation		
	p	$E(W)$	$P(W < 0.1)$	T	$E(W)$	$P(W < 0.1)$	T
$\lambda = 0.01$	0%	0.0000	1.0000	0.0011	0.0000	1.0000	0.0000
	25%	0.0126	0.9654	18.4991	0.0125	0.9662	18.5000
	50%	0.0250	0.9317	18.9983	0.0250	0.9323	19.0000
	75%	0.0376	0.8958	19.4987	0.0375	0.8985	19.5000
	100%	0.0502	0.8640	19.9991	0.0500	0.8647	20.0000
$\lambda = 0.1$	0%	0.0000	1.0000	0.0501	0.0000	1.0000	0.0000
	25%	0.0135	0.9621	17.3916	0.0125	0.9662	18.5000
	50%	0.0264	0.9264	18.3770	0.0250	0.9323	19.0000
	75%	0.0389	0.8930	19.0035	0.0375	0.8985	19.5000
	100%	0.0553	0.8574	18.2196	0.0500	0.8647	20.0000

Consider then a situation where $s-1$ agents work on outbound jobs. After the service completion of a new arrived call, the involved agent starts doing outbound jobs with a probability p or remains idle with a probability $1-p$. As a consequence, the probability that the system has s busy agents on outbound jobs is p and the probability that the system has $s-1$ busy agents on outbound jobs and one idle agent is $1-p$. Therefore, an arbitrary new call arriving to the system does not wait for service with probability p, and has to wait an exponential duration with rate μ_0 with probability $1-p$. This gives the proof of the expressions of the expected value and the cdf of W. Note that the result agrees with Equation (10) for the single server case. For $p > 0$, the system converges to a situation with $s-1$ agents working all the time on outbound jobs, and one agent that works on outbound jobs with probability p times the proportion of time during which this server does not work on inbound calls. The latter proportion is approximately 1. This finishes the proof of the throughput result, and also that of the proposition. □

In Table 4, we compare between the light-traffic approximation and simulation. We observe that the simulation results are close to the approximate ones for a very low workload. One can make use of the light-traffic approximation to address the routing optimization problem. Under the light-traffic regime, the presence of calls in the system can be neglected. The parameter q does not thus impact the results. The only parameter to focus on for Problem (1) is p. For a choice limited to the extreme cases, we should choose Model 4 if $\frac{1}{s\mu_0} \leq w^*$ (or $1 - e^{-\mu_0 s A_{WT}} \geq SL$). Otherwise Model 3 is the best. The optimal value of p with the constraint $P(W < A_{WT}) \geq SL$ is $p = SLe^{s\mu_0 A_{WT}}$. The optimal expected throughput is then $(s-1)\mu_0 + \mu_0 SLe^{s\mu_0 A_{WT}}$. The optimal value of p with a the constraint $E(W) \leq w^*$ is $p = s\mu_0 w^*$. The optimal expected throughput is then $(s-1)\mu_0 + s\mu_0^2 w^*$.

Heavy-traffic Approximation. Proposition 4 provides the approximate performance measures under a heavy-traffic regime.

Proposition 4 *Let* $\frac{1}{\mu_{eq}} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{q}{\mu_0} + \frac{1}{\mu_3}$. *We have*

$$E(W) \underset{\lambda \to s\mu_{eq}}{\sim} \left[\sum_{x=0}^{s-1} \frac{(\lambda/\mu_{eq})^x}{x!} + \frac{\frac{(\lambda/\mu_{eq})^s}{s!}}{1-\lambda/s\mu_{eq}} \right]^{-1} \frac{(\lambda/\mu_{eq})^s}{2\mu_{eq}(s-1)!(s-\lambda/\mu_{eq})^2} \tag{23}$$

$$\left(1 + \mu_{eq}^2 \left(\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{2q-q^2}{\mu_0^2} + \frac{1}{\mu_3^2} \right) \right), \tag{24}$$

$$T \underset{\lambda \to s\mu_{eq}}{\sim} sq \left(\frac{1}{\mu_0} + \frac{1}{\mu_2} \right) \mu_{eq}, \tag{25}$$

and

$$P(W < A_{WT}) \underset{\lambda \to s\mu_{eq}}{\sim} 1 - e^{-\frac{2s\left(1-\frac{\lambda}{s\mu_{eq}q}\right)\mu_{eq}A_{WT}}{1+\left(\frac{1}{\mu_1^2}+\frac{1}{\mu_2^2}+\frac{2q-q^2}{\mu_0^2}+\frac{1}{\mu_3^2}\right)\mu_{eq}^2}}. \tag{26}$$

Proof. Under a heavy-traffic regime, there is always at least one inbound call waiting in the queue. Since inbound calls have a non-preemptive priority over outbound jobs, there would not be then a possibility to route outbound jobs between calls. Outbound jobs are only treated inside an inbound call conversation. The system can be therefore approximated by an M/G/s queue with mean arrival rate λ and expected service rate $\frac{1}{\mu_{eq}}$. We next use the Lee and Longton (1959) approximation (see also Whitt (1983)); $E(M/G/s) = E(M/M/s) \times \frac{1+c_v^2}{2}$, where c_v^2 in our case is given by $c_v^2 = \mu_{eq}^2 \left(\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{2q-q^2}{\mu_0^2} + \frac{1}{\mu_3^2} \right)$. This proves Equation (23). Note that combining $s = 1$, $p = 0$ and this formula agrees with Equation (10). For the call waiting time cdf, Whitt (1983) shows that, for an M/G/s queue under a heavy-traffic regime, the distribution of $(1-\rho)W$ converges to an exponential distribution with rate $\frac{2\mu_{eq}}{1+c_v^2}$. This implies Equation (26). Since outbound jobs are only treated inside an inbound call conversation, the probability that an agent is working on outbound jobs is $\mu_{eq} \left(\frac{q}{\mu_0} + \frac{q}{\mu_2} \right)$. So, the outbound job expected throughput is $sq \left(\frac{1}{\mu_0} + \frac{1}{\mu_2} \right) \mu_{eq}$. This finishes the proof of the proposition. □

In Table 5, we compare between the heavy-traffic approximation and simulation. We observe that the simulation results converge to the approximate ones as the workload increases (q increases). The only parameter here is q. A simple analytical analysis, as that under a light-traffic regime, is not possible here. One can then numerically optimize the parameter q, using Equations (23)-(26). For a choice limited to the extreme cases, as the workload increases, we should first choose Model 4 then Model 2.

Table 5: Light-traffic approximation ($\lambda = 3.8$, $\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $s = 10$, $p = 50\%$)

	Simulation			Approximation		
q	$E(W)$	$P(W < 0.1)$	T	$E(W)$	$P(W < 0.1)$	T
0%	0.7614	0.0979	1.3148	0.8999	0.0676	1.4286
25%	1.8451	0.0403	1.7519	1.9836	0.0383	1.8644
50%	8.8221	0.0117	2.1992	8.9706	0.0104	2.2581
55%	19.4514	0.0054	2.3314	19.4617	0.0050	2.3323
58%	56.3273	0.0020	2.3757	56.3274	0.0018	2.3761
59.50%	631.1100	0.0069	2.3976	631.1100	0.0069	2.3978

6 Extensions

Using simulation, we assess to what extent the results of the previous sections still apply for more general settings closer to real-life call centers. In particular, we check wether the result, which states that at least one of the control parameters should have an extreme value, still hold or not. The modeling is generalized by considering non-Markovian service time distributions, non-stationary arrival process, and customer abandonment from the queue. In Sections 6.1 and 6.2, we describe the simulation experiments of the new settings and discuss the related observations, respectively.

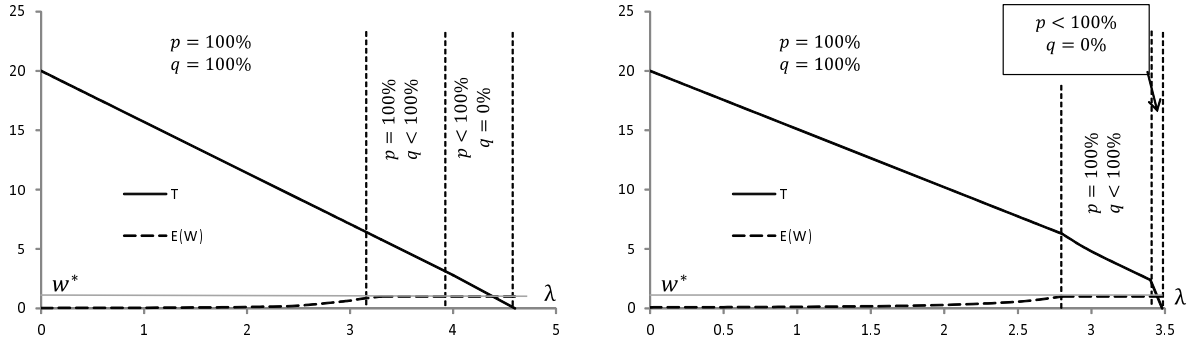
6.1 Simulation Experiments

In the simulation experiments, we only include one new future each time. This is to isolate the impact of each feature, which allows for a better understanding of the results.

Non-Markovian Service Phase Durations. Some studies have compared empirical distributions of service durations to exponential distributions and found an acceptable fit. One example is Kort (1983), which summarizes models of the Bell System Public Switched Telephone Network, developed in the 70's and 80's. Another example is that of Harris et al. (1987) for IRS call centers. More recent studies propose two other parametric statistical families that arise in applications: Erlang (or, more generally, Gamma) distribution and the lognormal distribution. Both families are explored in Chlebus (1997), who analyzes holding-time distributions in cellular communication systems. Other confirmations for the log-normal fit for call center service times are given in Bolotin (1994), Mandelbaum et al. (2000), and in Brown et al. (2005). Note also that mixtures of Erlang distributions are dense among all non-negative distributions. This sub-family of phase-type distributions can be then appropriately employed (Latouche and Ramaswami, 1999).

Since service time distribution may vary in practice, we assess the impact of the service time variability on the results for Problem (1). We choose two different variability levels ($c_v = 0$ and $c_v = 2$) instead of that of the exponential distribution ($c_v = 1$). We consider the deterministic (Figure 11(a)) and the log-normal distributions (Figure 11(b)) for the three service phases of inbound calls.

In both distribution cases, we use the same parameter values as those used in Figure 10(a) so as to have a coherent comparison. The ratio $1/\mu_i$ represents the expected length of service phase i ($i = 1, 2, 3$). In Figure 11(b), the standard deviation of each service phase is chosen such that it doubles the expected length of the phase.



(a) Deterministic service phase distributions ($c_v = 0$) (b) Log-normal service phase distributions ($c_v = 2$)

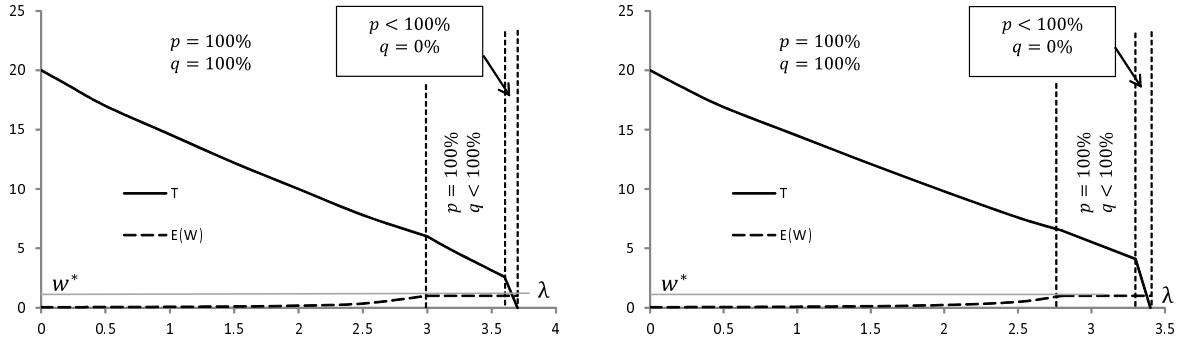
Figure 11: Optimal p and q ($\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $w^* = 1$, $s = 10$)

Time-Dependent Arrival Process. In real-life call centers, the mean arrival rate of calls is not constant but time-dependent (Akşin et al., 2007). We refer the reader to the recent survey of Whitt (2016) for the analysis of queues with time-dependent arrival rates. A common appropriate assumption for the process of arrivals is to consider a non-homogeneous Poisson process (Kim and Whitt, 2014). Following Ibrahim and Whitt (2011); Jouini et al. (2015), we propose here a simulation model with a non-homogeneous Poisson process where the arrival rate follows a sinusoidal function of the time

$$\lambda(t) = \lambda + a \sin(f \cdot t),$$

where λ is the average arrival rate, a is the amplitude, and f is the frequency. In order to provide an insightful illustration, we use the numerical values of Figure 10(a). To avoid negative values for the arrival rate, we choose $a = 0.5\lambda$ in Figure 12(a) and $a = 0.8\lambda$ in Figure 12(b). Therefore, both the average and the amplitude of the arrival rate increase in λ.

Customers Abandonment. An important feature in call centers is abandonment. We extend here the modeling by allowing inbounds to be impatient. After entering the queue, an inbound call will wait a random length of time for service to begin. If service has not begun by this time the call will abandon and be lost. We assume here that the abandonment time of inbounds from the queue is exponentially distributed with rate γ. Call center managers are concerned about the proportion of abandonments of inbounds. We update the optimization problem by considering a service level constraint on abandonments instead of that on waiting times. The optimization problem thus



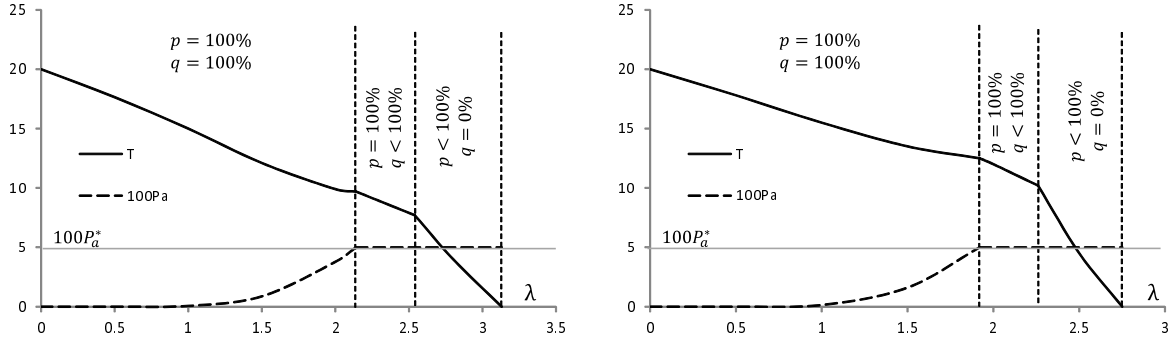
(a) $a = 0.5\lambda$ (b) $a = 0.8\lambda$

Figure 12: Optimal p and q ($\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $w^* = 1$, $s = 10$, $f = 0.1$)

becomes

$$\begin{cases} \text{Maximize the expected throughput of outbound jobs} \\ \text{subject to a service level constraint on the proportion of abandonments.} \end{cases} \tag{27}$$

In Figure 13, we answer Problem (27) for the same parameter values as those in Figure 10(a). We denote by P_a the proportion of abandonments. A common service level objective in practice is that less than 5% of calls abandon the queue (Akşin et al., 2007). We denote this threshold by P_a^*, $P_a^* = 5\%$.



(a) $\gamma = 1$ (b) $\gamma = 10$

Figure 13: Optimal p and q ($\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $s = 10$, $P_a^* = 5\%$)

6.2 Discussion

From the simulation experiments for the general settings, the conclusion which states that at least one of the two control parameters should be extreme still hold in all cases. Moreover:

- The maximal value for λ for which it is possible to answer Problem (1) or Problem (27) decreases with the variability of the service time (Figure 11), the amplitude of the arrival rate (Figure 12) or the abandonment rate (Figure 13). The reason is related to the increasing

difficulty to meet the service level constraint (performance deterioration) as the variability in service or arrival increase, or as the abandonment rate increases.

- The expected waiting time is more affected than the throughput of outbound jobs by the changes in the service time distribution or the arrival process. This can be seen through the curves of the throughput which are very similar in Figures 10(a), 11(a), 11(b), 12(a) and 12(b), whereas the curves of the expected waiting time increase faster in λ as the variability in service or arrival processes increase. In the single server case, these observations may be explained by the similarity between our model and an M/GI/1 queue. It has been shown (Chapter 5 in Kleinrock (1975)) that in an M/GI/1 queue, the probability of an empty system is only function of the first moment of the service time whereas the expected waiting time is function of its two first moments. The justification of this observation for the time-dependent arrival process is similar to the one for the variability of the service times by considering the GI/M/1 queue instead of the M/GI/1 queue (Chapter 6 in Kleinrock (1975)).

- In Figure 13, we observe that the same qualitative conclusions hold when defining the service level objective on the proportion of abandonment (instead of waiting time). Moreover, we observe that the use of the break reduces with customer impatience.

In summary, we observe that the conclusions from the analysis of the stylized modeling still hold qualitatively for the more general real-life modeling. Moreover, the applicability of the results is supported by previous findings in the literature: (i) it has been shown that the variability of service times is not important for large call centers. The performance mainly depend on their expected value (Mandelbaum and Schwartz, 2002; Whitt, 2005); (ii) the time-dependent arrival process can be approximated, as usually done, by considering multiple short-interval stationary parameters. It has been shown in the literature (Gans et al., 2003; Brown et al., 2005) that it is appropriate to assume constant parameter values during short intervals of the working day of 30 or 60 minutes. In these intervals, the stationary regime is reached and therefore the results may be applied interval by interval.

7 Conclusions

We considered a blended call center with inbound calls and outbound jobs. The call service is characterized by successive stages where one of them is a break for the agent in the sense that inside the conversation there is no required interaction during a non-negligible time between the two parties. We addressed an important question in the call center practice: how should managers make use of this opportunity in order to better improve performance? We focused on the optimization of the outbound job routing given that calls have a non-preemptive priority over outbound jobs.

Our objective was to maximize the expected throughput of outbound jobs subject to a constraint on the call waiting time.

We developed a general framework (Model PM) with two probabilistic parameters for the outbound job routing to agents. One parameter controls the routing between calls, and the other does the control inside a call conversation. We have also considered four particular cases corresponding to the extreme values of the probabilistic parameters. We derived various structural results for the single server case. We have also numerically illustrated and discussed the theoretical results in order to provide guidelines to call center managers. In particular, we proved for the optimal routing that all the time at least one of the two outbound job routing parameters has an extreme value. We then focused on the routing optimization problem for the multi-server case and considered a more general setting, using simulation and approximations developed under the light and heavy-traffic regimes. We found that most of the observations of the single server case are still valid (in particular the result stating that at least one control parameter has an extreme value).

There are several avenues for future research. It would be interesting to extend the structural results to the multi-server case. It would also be useful but challenging to extend the analysis to cases with an additional channel, in particular the chat which is increasingly used in call centers. Using the chat channel, an agent may handle many customers at the same time, which represent an additional opportunity to efficiently use the agent time.

Acknowledgements

We want to express our gratitude to Sébastien Thorel from INTERACTIV GROUP for the useful discussions. We also want to thank two anonymous reviewers and the editor for their comments, that significantly improved the paper.

References

Akşin, O., Armony, M., and Mehrotra, V. (2007). The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688.

Amari, S. and Misra, R. (1997). Closed-form expressions for distribution of sum of exponential random variables. *IEEE Transactions on Reliability*, 46(4):519–522.

Armony, M. and Maglaras, C. (2004). Contact centers with a call-back option and real-time delay information. *Operations Research*, 52(4):527–545.

Bernett, H., Fischer, M., and Masi, D. (2002). Blended call center performance analysis. *IT Professional*, 4(2):33–38.

Bhulai, S. and Koole, G. (2003). A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, 48(8):1434–1438.

Bolotin, V. (1994). Telephone circuit holding time distributions. *Proceedings of the 14th International Teletraffic Conference. Labetoulle J.and Roberts J.W., editors*, pages 125–134.

Brandt, A. and Brandt, M. (1999). A two-queue priority system with impatience and its application to a call center. *Methodology and Computing in Applied Probability*, 1:191–210.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50.

Charron, S. and Koechlin, E. (2010). Divided Representation of Concurrent Goals in the Huma Frontal Lobes. *Science*, 328:360–363.

Chlebus, E. (1997). Empirical validation of call holding time distribution in cellular communications systems. In *Proceedings of the 15th International Teletraffic Conference*, volume 6, page 2.

Choudhury, G., Lucantoni, D., and Whitt, W. (1995). Numerical solution of piecewise-stationary of Mt/Gt/1 queues. *Operations Research*, 45(3):451–463.

Daigle, J. and Lucantoni, D. (1991). Queueing systems having phase-dependant arrival and service rates. Chapter 10 of Numerical Solutions of Markov Chains, Editor: W.J. Stewart, Marcel Dekker, INC., 161-202.

Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., and Avramidis, A. (2007). Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645.

Dux, P., Tombu, M., Harrison, S., Rogers, B., Tong, F., and Marois, R. (2009). Training improves Multitasking Performance by Increasing the Speed of Information Precessing in Human Prefontal Cortex. *Neuron*, 63:127–138.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):73–141.

Gans, N. and Zhou, Y. (2003). A call-routing problem with service-level constraints. *Operations Research*, 51(2):255–271.

Gevros, P., Crowcroft, J., Kirstein, P., and Bhatti, S. (2001). Congestion control mechanisms and the best effort service model. *IEEE Network*, 15(3):16–26.

Gladstones, W., Regan, M., and Lee, R. (1989). Division of attention: The single-channel hypothesis revisited. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 41(1):1–17.

Gourdon, X. (1994). *Les maths en tête : Algèbre*. Ellipses, Paris.

Guo, P. and Zipkin, P. (2008). The effects of information on a queue with balking and phase-type service times. *Naval Research Logistics*, 55(5):406–411.

Harris, C., Hoffman, K., and Saunders, P. (1987). Modeling the irs telephone taxpayer information system. *Operations Research*, 35(4):504–523.

Ibrahim, R. and Whitt, W. (2011). Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, 59(5):1106–1118.

Jouini, O., Akşin, O. Z., Karaesmen, F., Aguir, M. S., and Dallery, Y. (2015). Call center delay announcement using a newsvendor-like performance criterion. *Production and Operations Management*, 24(4):587–604.

Keblis, M. and Chen, M. (2006). Improving customer service operations at amazon.com. *Interfaces*, 36(5):433–445.

Keilson, J., Sumita, U., and Zachmann, M. (1987). Row-continuous finite markov chains: Structure and algorithms. *Journal of the Operations Research Society of Japan*, 30(3):291–314.

Kim, S. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480.

Kleinrock, L. (1975). *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication.

Kort, B. (1983). Models and methods for evaluating customer acceptance of telephone connections. *Proc. IEEE GLOBECOM'83*, pages 706–714.

Latouche, G. and Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. SIAM.

Le Boudec, J. (1998). Application of network calculus to guaranteed service networks. *IEEE Transactions on Information theory*, 44(3):1087–1096.

Lee, A. and Longton, P. (1959). Queueing process associated with airline passenger check-in. *Operational Research Quarterly*, 10:56–71.

Legros, B. and Jouini, O. (2015). A linear algebraic approach for the computation of sums of erlang random variables. *Applied Mathematical Modelling*, 39(16):4971–4977.

Legros, B., Jouini, O., and Koole, G. (2015). Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 47(4):414–430.

Legros, B., Jouini, O., and Koole, G. (2016). Optimal scheduling in call centers with a callback option. *Performance Evaluation*, 95:1–40.

Mandelbaum, A., Sakov, A., and Zeltyn, S. (2000). Empirical analysis of a call center. *URL http://iew3. technion. ac. il/serveng/References/ccdata. pdf. Technical Report*.

Mandelbaum, A. and Schwartz, R. (2002). Simulation experiments with M/G/100 queues in the Halfin-Whitt (QED) regime. *Technical Report*.

Mitrani, I. and Chakka, R. (1995). Spectral Expansian Solution of a Class of Markov Models: Application and Comparison with the Matrix-Geometric Method. *Performance Evaluation*, 23:241–260.

Neuts, M. (1995). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Dover Publications, Revised edition.

Pang, G. and Perry, O. (2014). A logarithmic safety staffing rule for contact centers with call blending. *Management Science*, 61(1):73–91.

Pichitlamken, J., A., D., P., L., and Avramidis, A. (2003). Modeling and simulation of a telephone call center. *Proceedings of the 37th Conference on Winter Simulation, New Orleans, LA*, pages 1805–1812.

Pollaczeck, F. (1930). Uber eine Aufgabe der Wahrscheinlichkeitstheorie. *Mathematische Zeitschrift*, 32:64–100.

Seelen, L. (1986). An algorithm for Ph/Ph/c queues. *European Journal of Operational Reseearch*, 23(1):118–127.

Whitt, W. (1983). Comparison conjectures about the M/G/s queue. *Operations Research Letters*, 2(5):203–209.

Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235.

Whitt, W. (2016). Queues with time-varying arrival rates: A bibliography.

Zhang, H. (1995). Service disciplines for guaranteed performance service in packet-switching networks. *Proceedings of the IEEE*, 83(10):1374–1396.

Benjamin Legros is Professor in operations management at EM Normandie. He received a B.Sc. degree in industrial engineering from CentraleSupélec in 2006. He carried out his Ph.D. research on the optimization of multi-channel call centers at CentraleSupélec and received a Ph.D degree in 2013. His current research interests are in stochastic modeling, queueing theory and operations management of call centers.

Oualid Jouini is Professor in operations management at CentraleSupélec. He received a B.Sc. degree in Industrial Engineering from Ecole Nationale d'Ingénieurs de Tunis in 2001 and a M.Sc. degree in Industrial Engineering from CentraleSupélec in 2003. He received a Ph.D. degree on the optimization of call centers at CentraleSupélec in 2006 and held a postdoc position at University of Minnesota in 2007. He holds the chair *Call Centers* at CentraleSupélec. His current research interests are in stochastic modeling and service operations management. His main application areas are call centers and healthcare systems.

Ger Koole Dr. Ger M. Koole is Professor at VU University Amsterdam. He graduated at Leiden University in 1992 on a thesis on the control of queueing systems and held post-doc positions at CWI Amsterdam and INRIA Sophia Antipolis. He holds the chair "optimization of business processes" which makes him responsible for research in applied operations research and for the bachelor and master programs in business analytics.

He has supervised 14 PhD students and published over 90 papers in the international literature. Next to his academic work Dr. Koole co-founded three companies: the call center planning company CCmath, the internet advertisement company Adscience, and the hotel revenue management company IrevenU. He is also a founder of PICA, the VU university/medical center joint knowledge center on health care operations management, and of ACBA, the multi-disciplinary "Amsterdam Center for Business Analytics" which has been very successful in obtaining company funding for academic research.