# Optimal scheduling in call centers with a callback option

Benjamin Legros, Oualid Jouini, Ger Koole

# Optimal Scheduling in Call Centers with a Callback Option

Benjamin Legros[1] • Oualid Jouini[1] • Ger Koole[2]

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

[2] *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

***Performance Evaluation***. To appear, 2015.

### Abstract

We consider a call center model with a callback option, which allows to transform an inbound call into an outbound one. A delayed call, with a long anticipated waiting time, receives the option to be called back. We assume a probabilistic customer reaction to the callback offer (option). The objective of the system manager is to characterize the optimal call scheduling that minimizes the expected waiting and abandonment costs. For the single-server case, we prove that non-idling is optimal. Using a Markov decision process approach, we prove for the two-server case that a threshold policy on the number of queued outbound calls is optimal. For the multi-server case, we numerically characterize a switching curve of the number of agents reserved for inbound calls. It is a function of the number of queued outbound calls, the number of busy agents and the identity of jobs in service. We also develop a Markov chain method to evaluate the system performance measures under the optimal policy.

We next conduct a numerical study to examine the impact of the policy parameters on the system performance. We observe that the value of the callback offer is especially important for congested situations. It also appears that the benefits of a reservation policy are more apparent in large call centers, while they almost disappear in the extreme situations of light or heavy workloads. We moreover observe in most cases that the callback offer should be given upon arrival to any delayed call. However, if balking and abandonment are very high (which helps to reduce the workload) or if the overall treatment time spent to serve an outbound call is too large compared to that of an inbound one, there is a value in delaying the proposition of the callback offer.

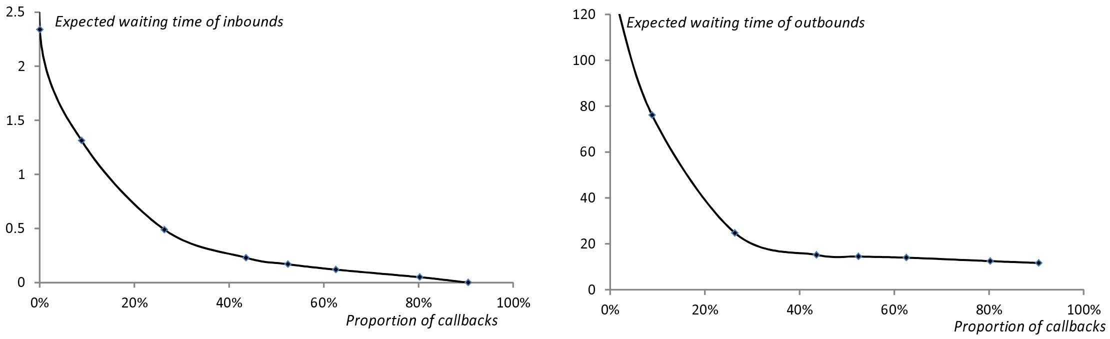
**Keywords.** Call centers, callback option, routing optimization, queueing systems, Markov chains, Markov decision processes, switching curve, reservation policy, blending operations, performance measures.

## 1 Introduction

**Context and Motivation.** Call centers serve as the public face in various areas and industries: insurance companies, emergency centers, banks, information centers, help-desks, tele-marketing, just to name a few. The success of call centers is due to the technological advances in information and communication systems. The most used form of communication is the telephone. However, in the context of highly congested call centers, the use of alternative service channels can be proposed to customers so as to better match demand

and capacity. Alternative channels could be email, chat, blog, or postponed callback service. We focus on this last alternative. The idea is that customers, who are expected to experience long waiting times, receive the option to be called back later. This leads to a contact center with two channels, one for inbound calls (inbounds), and another for outbound calls (outbounds). The recent study of ICMI (2013), based on the analysis of 361 large contact centers, reports that 76% of them use the outbound channel.

The flexibility of the callback option comes from the willingness of some customers to accept future processing. The call center can then make use of this opportunity to better manage arrival uncertainty, which in turn would improve the system performance. An illustration of callback option benefits is provided in Figure 1. The figure gives simulated performance measures of a call center example with various levels for the use of the callback option. We consider a non-idling system where inbounds have a non-preemptive higher priority over outbounds. We observe that the expected waiting times of inbound and outbound calls are considerably improved by using the callback option. For instance, the expected waiting time of inbounds could be divided by around 20 (it decreases from 8 minutes and 55 seconds to 23 seconds) while only 10% of arriving calls choose to be called back.



(a) Inbounds

(b) Outbounds

Figure 1: Effect of the callback option on performance (arrival rate = 5.5, service rate = 0.2, number of agents = 28)

The unpredicted and flexible call center environment offers the potential for a routing optimization that would lead to a significant operational improvement. It is a non-expensive approach compared to staffing optimization (Gans and Zhou, 2003; Akşin et al., 2007). One important question for managers in our context is how should be the routing rule of jobs that would ensure non-excessive waiting times for both job types, i.e., upon a service completion, should the agent handle an inbound or an outbound call? when should be proposed the callback offer? We address these questions under a queueing modeling framework and a probabilistic customer reaction to the callback option.

A call center where agents simultaneously handle inbound and outbound calls is commonly referred to as *call blending*. The key distinction of call center problems with blending comes from the fact that outbound tasks have less urgency relative to inbound calls. Blended operations problems have led to research on performance evaluation (Bernett et al., 2002; Pichitlamken et al., 2003; Deslauriers et al., 2007), staffing (Pang and Perry, 2014) and analysis of blending policies (Gans et al., 2003; Bhulai and Koole, 2003; Armony and Maglaras, 2004a; Armony and Ward, 2010; Legros et al., 2013, 2015b). Because of the lack of service level requirement on outbounds, it is best to give higher priority to inbounds. Moreover, to reduce the number of inbounds who may experience long waiting before service, one has to guarantee that there is sufficient idleness in the system. In the patent of Dumas et al. (1996), based on extensive simulation experiments, it is shown that blending inbound and outbound calls and employing a threshold policy, ensure that the outbound throughput rate is met while waiting times of inbounds are very short. It is also shown that blending the two types of calls in one pool requires less agents than employing two distinct pools. Bhulai and Koole (2003) and Gans and Zhou (2003), prove this optimal control, which is of threshold type, when the service rates of the two types of jobs are equal. More precisely, they show that it is optimal to schedule outbound tasks only when no outbounds are in the queue and the number of idle agents exceeds a certain threshold.

In the case of a callback option, this policy can not be directly applied. The reason is that the above literature considers an infinite amount of non-priority jobs. In a call center with a callback option, the number of customers waiting to be called back has to be finite in order to avoid infinite waiting. The routing policy should then account for the length of the callback queue. Another difference, compared to cases with classical infinite amount of outbound tasks, is that inbound and outbound arrivals are negatively correlated. This requires further analysis, and may lead to different managerial recommendations.

**Contributions.** We consider a call center with a single customer type. A delayed call, with a long anticipated waiting time, receives the option to be called back. We develop a modeling that accounts for balking, abandonment, probabilistic customer reaction to a state-dependent delay information, unequal service requirements for job types, and the eventual non-availability of a called back customer. The objective of the system manager is to find the optimal call scheduling policy that minimizes the expected operating costs of inbounds and outbounds. The control actions concern the number of agents reserved for inbounds and the system state situations at which the callback offer should be proposed.

We distinguish three main contributions. The first contribution is related to the agent reservation policy. We prove for the single-server case that non-idling is optimal. Using a Markov decision process (MDP) approach, we prove for the two-server case with equal service requirements that a threshold policy on the number of queued outbounds is optimal. Based on the two-server result, we conjecture for the multi-server case that the optimal policy is of switch type. The number of agents to reserve for inbounds depends on the number of queued outbounds, the number of busy agents and the identity of jobs in service. Moreover, we examine the impact of the system exogenous parameters on the agent reservation policy. We observe, for example, that a reservation policy is not likely to be used under light or heavily loaded situations.

The second contribution is the performance analysis under the optimal reservation policy. The performance measures of interest are related to the job type waiting times and abandonments. We develop a controlled numerical approximation to obtain these performance measures for the general modeling. For various particular cases, using a Markov chain method, we go further by providing either exact numerical algorithms, or closed-form expressions for the performance analysis.

The third contribution is the analysis of the impact of the policy parameters on performance. We derive the first and second monotonicity results in the number of agents for the performance measures in the non-idling case. These results support that the benefit of a reservation policy is more apparent in large call centers. Moreover, in most cases, the callback offer should be given upon arrival to any delayed call. We prove this result in the non-idling case using first order monotonicity results. However, if balking and abandonment are very high (which helps to reduce the workload) or if the overall treatment time spent to serve an outbound call is too large compared to that of an inbound one, there is a value in delaying the callback offer to all customers.

**Literature Review.** There is a rich literature on the operations management in call centers. We refer the reader to the two surveys by Gans et al. (2003) and Akşin et al. (2007). For a background on the specific context of multi-channel call centers, we refer the reader to Chapter 7 in Koole (2013).

As mentioned above, there are only few papers dealing with routing strategies in the context of a finite amount of callbacks. The first two papers directly addressing the problem of the callback option are by Armony and Maglaras (2004a,b). The authors consider a model in which customers are given a choice of whether to wait online for their call to be answered or to leave a number and be called back within a specified time or to immediately balk. Upon arrival, customers are informed (or know from prior experience) of the expected waiting time if they choose to wait and the delay guarantee for the callback option. Their decision is probabilistic and based on this information.

Under the heavy-traffic regime, Armony and Maglaras (2004a) develop an estimation scheme for the anticipated real-time delay. They also propose an asymptotically optimal routing policy that minimizes real-time delay subject to a deadline on the postponed service mode. In Armony and Maglaras (2004b), the authors develop an asymptotically optimal routing rule, characterize the unique equilibrium regime of the system, and propose a staffing rule that picks the minimum number of agents that satisfies a set of operational constraints on the performance of the system. To the contrary to Armony and Maglaras (2004a,b), we account here for the feature of abandonment, unequal service requirements and the possible non-availability of an outbound call. Yet, our modeling is restricted to policies with strict non-preemptive priority for inbounds. Armony and Maglaras (2004a,b) consider instead a state-dependent priority policy.

Two recent papers are by Kim et al. (2012) and Dudin et al. (2013). Kim et al. (2012) consider a call center model with a callback option where the queue capacity for inbounds is finite. As in our modeling, customer balking and abandonment are allowed. The authors provide an efficient algorithm for calculating the stationary probabilities of the system states. Moreover, they derive the Laplace-Stieltjes transform of the sojourn time distribution of virtual customers. Dudin et al. (2013) consider a slightly different modeling, where lost customers are called back. There are two agent teams, one that handles in priority inbounds, and another one that handles in priority outbounds. They compute the stationary probabilities, and deduce the system performance measures. They also numerically address the staffing issue for the two teams.

Our approach differs from those in Armony and Maglaras (2004a,b); Kim et al. (2012); Dudin et al. (2013) since we allow for agent reservation strategies. We also allow to control the proposition of the callback offer, whereas in all above references this option is proposed to all customers. Other papers considering finite amounts of outbound tasks are Armony and Ward (2010) and Gurvich et al. (2009). They study call centers that exercise cross-selling. The cross-selling phase is initiated by the agent and can thus be considered as a type of outbound work in finite number. However, these are less related to our specific context of callbacks.

**Structure of the paper.** The remainder of this paper is structured as follows. In Section 2, we describe the call center model with a callback option. In Section 3, we address the optimal routing problem for outbound calls. In Section 4, we evaluate the performance measures under the optimal reservation policy. In Section 5, we use the optimization and performance measures results to examine the impact of the policy parameters on performance. We then provide conclusions and highlight future research directions. Part of the proofs of the results of the main paper are given in the appendices and the online supplement.

# 2 Model Description

We consider a call center modeled as a multi-server queueing system with $s$ identical, parallel servers (agents). The call center handles two types of jobs: inbound calls (type 1 jobs or inbounds) initiated by customers, and outbound calls (type 2 jobs or outbounds) initiated by agents. Each agent can handle both types of jobs. Type 1 jobs request for a real-time service, while type 2 jobs are customers with a postponed service. A job 2 customer is originally a job 1 customer that has chosen to be called back. The real-time service is more important in the sense that the waiting time of an inbound call should be in the order of seconds or minutes, whereas the postponed service could be delayed for several hours. This is the attractive aspect for using the callback option. It allows to create a flexibility by delaying some of the workload for future processing, which would improve the system performance.

The arrival process of inbounds is assumed to be a homogeneous Poisson process with rate $\lambda$. Inbound calls arrive at a dedicated first come, first served (FCFS) queue with infinite capacity, denoted by queue 1. We assume that the service times for inbounds are i.i.d. and exponentially distributed with rate $\mu_1$. Customers in queue 1 can be impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time, the customer will abandon. Times before abandonment for inbounds are assumed to be i.i.d. and exponentially distributed with rate $\beta$. Because of the flexibility of type 2 jobs, the system manager allocates more capacity to real-time service. Type 1 jobs have therefore a strict non-preemptive priority over type 2 jobs, which means that if an agent is busy with a job 2, the agent will finish first this job before turning to a newly arrived job 1. The non-preemption priority rule is coherent with the common call center practice, where it is not appropriate to interrupt a conversation with a low priority customer. In addition, we allow for agent reservation policies for inbounds. In other words, we allow an agent to remain idle when queue 1 is empty and queue 2 is not. This may reduce the waiting time of future inbound arrivals. For similar multi-channel call center situations, agent reservation policies have been shown to be efficient (Bhulai and Koole, 2003; Legros et al., 2013).

If a customer accepts to be called back, she virtually joins a FCFS queue, denoted by queue 2. Due to the nature of the outbound demand, we consider for this customer, the three possibilities as follows. With probability $r_1$, she has exactly the same need as the one she had when she first made her call. In this case, the service time is assumed to be exponentially distributed with rate $\mu_1$ (similarly to an inbound customer). With probability $r_2$ ($r_1 + r_2 > 0$), she has already resolved her problem or a part of it. Hence, her service time may be shorter. We assume in this case that the service time is exponentially distributed with rate $\mu_2$ ($\mu_2 \geq \mu_1$). Finally, with the remaining probability $1 - r_1 - r_2$, the outbound customer is not available, and an agent will try again to call her back later on. To handle such a situation, we assume that the agent spends a random duration assumed to be exponentially distributed with rate $\mu_3$. This duration corresponds to the required time to leave a message to the customer, and to place her back in the queue at the last position (she will be called back when she will again reach the first position under the FCFS rule).

**Description of the call back option.** The state of the system at a given time $t$ is defined by four variables: $x$, $y$, $s_2$, $s_3$, where $x$ is the number of inbounds in queue 1 or in service plus the number of outbounds in service with the same service time requirement as inbounds (service rate $\mu_1$), $y$ is the number of outbounds in queue 2, $s_2$ is the number of agents busy with outbounds that require a fast service (service rate $\mu_2$), and $s_3$ is the number of agents handling non-available outbound situations (rate $\mu_3$), for $x, y \geq 0$ and $0 \leq s_2, s_3 \leq s$.

Consider a newly arriving inbound call. If at least one agent is available, the customer immediately starts service. If all agents are busy and the number of waiting calls in queue 1 is strictly lower than a given threshold, denoted by $k \in \mathbb{N}$, a delay information is announced to the customer. The delay information is based on the system state. We do not restrict the model to a specific type of information: it could be the length of queue 1, the expected value or some quantiles of the waiting time, etc. The new inbound customer then reacts to the delay information. She either balks (immediately leaves the system) with probability $\alpha_{x,s_2,s_3}$, or joins queue 1 with probability $1 - \alpha_{x,s_2,s_3}$ where she may abandon or start service after some time duration. We assume that the probability $\alpha_{x,s_2,s_3}$ increases in the announced delay, i.e., $\alpha_{x+1,s_2,s_3} \geq \alpha_{x,s_2,s_3}$, for $s \leq x + s_2 + s_3 < s + k$, $0 \leq s_2, s_3 \leq s$. Note that the probability $\alpha_{x,s_2,s_3}$ could be chosen constant for the case with no delay information.

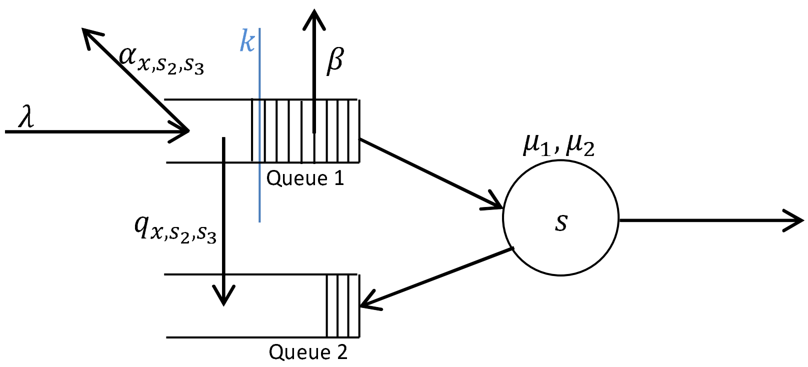


Figure 2: The callback option model

If the number of waiting calls in queue 1 is higher than or equal to $k$, the system provides a delay information as well as a callback option. Exceeding the threshold $k$ captures the fact that customers are likely to experience too long waiting times in case they would request for a real-time service. The delay information is system state-dependent. Concretely, the new inbound customer have the following three possibilities upon her arrival: she balks (immediately leaves the system) with probability $\alpha_{x,s_2,s_3}$, or she chooses the callback option and virtually joins queue 2 with probability $q_{x,s_2,s_3}$, or she joins queue 1 with probability $1 - q_{x,s_2,s_3} - \alpha_{x,s_2,s_3}$, for $x + s_2 + s_3 \geq s + k$, $0 \leq s_2, s_3 \leq s$. Again, we assume that $\alpha_{x+1,s_2,s_3} \geq \alpha_{x,s_2,s_3}$ and $q_{x+1,s_2,s_3} \geq q_{x,s_2,s_3}$ for $x + s_2 + s_3 \geq s + k$ and $0 \leq s_2, s_3 \leq s$. Also, the quantities $\alpha_{x,s_2,s_3}$ and $q_{x,s_2,s_3}$ could be chosen constant for $x + s_2 + s_3 \geq s + k$, $0 \leq s_2, s_3 \leq s$. In such a case, we will then simply write them as $\alpha$ or $q$ to simplify the presentation. An illustration of the model is given in Figure 2.

**Problem formulation.** Let us first define the performance measures of interest. We denote by $W_1$, $W_2$ and $W$ the random variables measuring the stationary waiting time of served inbounds in queue 1, the stationary waiting time of outbounds in queue 2, and the unconditional stationary waiting time in the queue of an arbitrary job (inbound or outbound), respectively. We also denote by $P_a$ the stationary proportion of inbounds that leave the system without service either by abandoning queue 1, or by balking upon arrival. The stationary proportion of inbounds that balk upon arrival is defined as $P_b$. We finally denote by $\psi$ the stationary probability that a new inbound call becomes an outbound one.

We consider an economic framework based on the holding costs of jobs 1 and 2, and the cost of lost calls (because of balking or abandonment). The objective of the system manager is to characterize the optimal routing policy which minimizes the expected system cost, denoted by $SC$, and given by

$$SC = \gamma_1 E(W_1) + \gamma_2 E(W_2) + \gamma_3 P_a,$$

where $\gamma_1$, $\gamma_2$ and $\gamma_3$ are the cost parameters, and where $E(Z)$ is the expected value of a given random variable $Z$. We assume that $\gamma_1 > \gamma_2$ to give more importance to the waiting time of inbounds than that of outbounds. The control parameters for the call center manager are the threshold $k$ for queue 1 which characterizes the callback option, and the agent reservation policy for inbounds.

For a given state $(x, y, s_2, s_3)$ ($0 \leq x + s_2 + s_3 < s$ and $y > 0$), there are two possible actions: the first one is to serve an outbound call and move to state $(x+1, y-1, s_2, s_3)$ with probability $r_1$, or to state $(x, y-1, s_2+1, s_3)$ with probability $r_2$, or to state $(x, y-1, s_2, s_3+1)$ with probability $1 - r_1 - r_2$; the second one is to keep the first outbound in line in queue 2 and stay at state $(x, y, s_2, s_3)$. The knowledge of the optimal actions at each state defines a function denoted by $c(x, y, s_2, s_3)$. The curve of this function separates the states where the optimal action is to serve an outbound call from those where it is optimal to keep an outbound call in queue 2. The function $c(x, y, s_2, s_3)$ defines therefore the agent reservation policy. It will be characterized in Section 3. A summary of the model notations is given in Table 1.

The call center model described above is referred to as *Model G* (general model). Because of its complexity, we define submodels that correspond to various special cases, for which it is easier to observe and prove insights. We denote by *Model A* the submodel where outbounds have the same service rate as inbounds and these are available when they are called back ($r_1 = 1$ and $r_2 = 0$), by *Model B* a submodel of Model A where inbounds are infinitely patient ($\beta = 0$), by *Model C* a particular case of Model B where the balking and callback parameters are assumed to be constant (for example when no information is given to arriving

Table 1: Model notations

| | |
|---|---|
| System state description | |
| $x$ | Number of inbounds in queue 1 or in service plus number of outbounds (with the same service requirement as inbounds) in service |
| $y$ | Number of outbounds in queue 2 |
| $s_2$ | Number of agents handling fast-served outbounds |
| $s_3$ | Number of agents handling non-available outbound situations |
| Exogenous parameters | |
| $\lambda$ | Arrival rate of inbounds |
| $s$ | Number of agents |
| $r_1$ | Probability that an outbound call has the same service requirement as an inbound one |
| $r_2$ | Probability that an outbound call has a shorter service requirement than an inbound one |
| $1 - r_1 - r_2$ | Probability that an outbound call in queue 2 is not available |
| $\mu_1$ | Service rate of inbounds, and also a part of outbounds with the same service requirement |
| $\mu_2$ | Service rate for fast-served outbounds |
| $\mu_3$ | Service rate for handling non-available outbounds |
| $\beta$ | Abandonment rate for each inbound call in queue 1 |
| $\alpha_{x,s_2,s_3}$ | Probability that a new inbound call balks upon arrival |
| $q_{x,s_2,s_3}$ | Probability that an inbound call accepts the callback offer upon arrival |
| Control parameters | |
| $k$ | Threshold on the length of queue 1, at which we start to propose the callback offer |
| $c(x, y, s_2, s_3)$ | Curve for the agent reservation policy |
| Performance Measures | |
| $\Psi$ | Proportion of inbounds that accept the callback offer |
| $P_a$ | Proportion of inbounds that leave the system without service (after a balking or an abandonment) |
| $E(W_1), E(W_2), E(W)$ | Expected waiting time for served inbounds in queue 1, expected waiting time for outbounds in queue 2, and unconditional waiting time in the queue of an arbitrary job (inbound or outbound), respectively |

customers). We also define *Model NI* (non-idling model) a submodel of Model G where idling is not allowed (i.e., the first outbound call in queue 2 starts service as soon as an agent becomes available and queue 1 is empty). An illustration of the submodels is depicted in Figure 3.

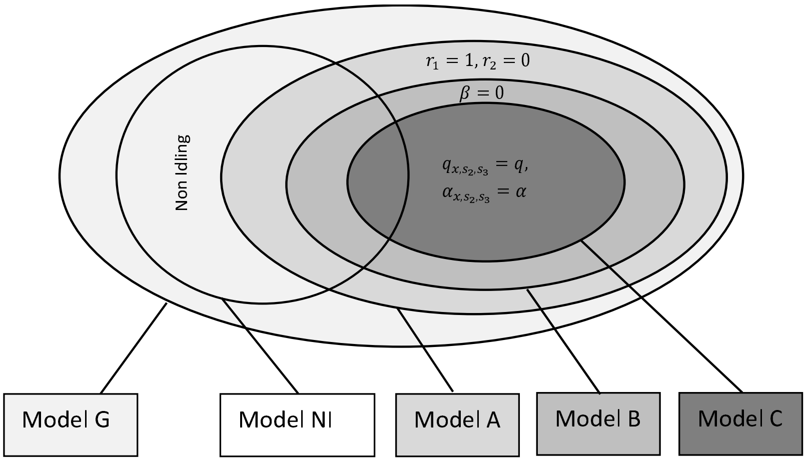


Figure 3: The submodels

**Markov decision process approach.** For Model G, we formulate the routing problem as a Markov decision process (MDP). Since we are considering long-term average performance, it is optimal to schedule jobs at arrival, service completion or abandonment times. If it is optimal to keep a server idle at a given time, then the action remains optimal until the next event in the system. This result follows directly from the continuous-time Bellman equation (Puterman (1994), Chapter 11). Therefore, it suffices to consider the system only at arrival, service completion or abandonment times. Due to the call abandonment in queue 1, the total event rate is not bounded. We therefore use the traditional approach where we assume that queue 1 has a limited capacity $N$ ($N \geq 0$). The parameter $N$ is chosen high enough to approximate the real system. The total event rate is then uniformly bounded by $\lambda + s \max(\mu_1, \mu_2, \mu_3) + N\beta$, and without loss of generality, we assume that it is equal to one. We next use the well known uniformization technique (Puterman (1994), Chapter 8), which allows to apply discrete-time dynamic programming to characterize the optimal routing policy.

The possible actions for an agent just after a service completion (and queue 1 is empty) are either to remain idle, or to serve an outbound call if queue 2 is not empty. We choose to formulate a 2-step value function, in order to separate transitions and actions and simplify the involved expressions. We define the sequences $U_n(x, y, s_2, s_3)$ and $V_n(x, y, s_2, s_3)$ over $n$ steps, for $n, x, y \geq 0$ and $0 \leq s_2, s_3 \leq s$. For $n \geq 0$, we have

$$
\begin{aligned}
&U_{n+1}(x, y, s_2, s_3) = \gamma_1(x + s_2 + s_3 - s)^+ + \gamma_2 y \qquad\qquad (1)\\
&+ \lambda \big[\mathbf{1}_{(0 \leq x+s_2+s_3<s)} V_n(x + 1, y, s_2, s_3)\\
&\qquad\qquad +\mathbf{1}_{(s \leq x+s_2+s_3<s+k)} \left((1 - \alpha_{x,s_2,s_3})V_n(x + 1, y, s_2, s_3) + \alpha_{x,s_2,s_3}(V_n(x, y, s_2, s_3) + \gamma_3)\right)\\
&\qquad\qquad + \mathbf{1}_{(s+k \leq x+s_2+s_3<s+N)}(q_{x,s_2,s_3} V_n(x, y + 1, s_2, s_3) + \alpha_{x,s_2,s_3}(V_n(x, y, s_2, s_3) + \gamma_3)\\
&\qquad\qquad + (1 - q_{x,s_2,s_3} - \alpha_{x,s_2,s_3})V_n(x + 1, y, s_2, s_3))\\
&\qquad\qquad +\mathbf{1}_{(x+s_2+s_3=s+N)}(q_{N-1,s_2,s_3} V_n(x, y + 1, s_2, s_3) + (1 - q_{N-1,s_2,s_3})(V_n(x, y, s_2, s_3) + \gamma_3))\big]\\
&+ \beta(x + s_2 + s_3 - s)^+(V_n(x - 1, y, s_2, s_3) + \gamma_3) + \min(s - s_2 - s_3, x)\mu_1 V_n(x - 1, y, s_2, s_3)\\
&+ s_2\mu_2 V_n(x, y, s_2 - 1, s_3) + s_3\mu_3 V_n(x, y + 1, s_2, s_3 - 1)\\
&+ \left(1 - \lambda - \beta(x + s_2 + s_3 - s)^+ - \min(s - s_2 - s_3, x)\mu_1 - s_2\mu_2 - s_3\mu_3\right) V_n(x, y, s_2, s_3), \text{ for } x, y \geq 0, \text{ and } 0 \leq s_2, s_3 \leq s,
\end{aligned}
$$

where $\mathbf{1}_{(x\in A)}$ is the indicator function of a subset $A$, and

$$
\begin{aligned}
&V_{n+1}(x, y, s_2, s_3)\\
&= \min(r_1 U_{n+1}(x + 1, y - 1, s_2, s_3) + r_2 U_{n+1}(x, y - 1, s_2 + 1, s_3) + (1 - r_1 - r_2)U_{n+1}(x, y - 1, s_2, s_3 + 1), U_{n+1}(x, y, s_2, s_3)),
\end{aligned}
$$

for $y > 0$ and $0 \leq x+s_2+s_3 < s$ and $V_{n+1}(x, y, s_2, s_3) = U_{n+1}(x, y, s_2, s_3)$ in the remaining cases. We choose $V_0(x, y, s_2, s_3) = U_0(x, y, s_2, s_3) = 0$, for $x, y \geq 0$, and $0 \leq s_2 + s_3 \leq s$. The transitions at boundary states $x + s_2 + s_3 = N$ are chosen such that the monotonicity properties of the value functions are maintained. The value of this choice is proven in the proof of Theorem 1 in Section 3.2. Another possibility to maintain the monotonicity properties is to use the smoothed rate truncation as proposed by Bhulai et al. (2014), however, this would imply a more complicate expression of the value functions in our setting.

The long-term average optimal actions can be obtained through value iteration, by recursively evaluating $V_n$ using Equation (1), for $n \geq 0$. As $n$ tends to infinity, the minimizing actions converge to the optimal ones (Puterman, 1994). For $0 \leq x + s_2 + s_3 < s$ and $y > 0$, the minimizing action is chosen between keeping an outbound call in queue 2 or starting the service of this call. For $x + s_2 + s_3 \geq s$, we do not consider any control action because of the priority for inbounds (i.e., no possibility of having an idle agent while a call is waiting in queue 1).

# 3 Optimal Agent Reservation Policy

We consider the single, the two-server and the multi-server cases. For the multi-server case of Model G, we first prove a preliminary result stating that when all agents are idling and queue 2 is not empty, then it is optimal to serve at least the first outbound call in line. A corollary of this result is that non-idling is optimal in the single-server case. In the two-server case, we prove in Theorem 1 the optimal reservation policy for Model A. It is a threshold policy on the number of waiting outbounds in queue 2. For the multi-server cases of Models A and G, we conjecture that the optimal routing follows a state-dependent threshold policy, i.e., a switching curve. For Model A, the switching curve is only based on the number of outbounds in queue 2 and the number of busy agents. In addition to that, for Model G, the optimal policy depends on the number of each job type in service.

The result for the multi-server case is intuitive and a standard extension, in MDP problems, of the proved result in the single and two-server cases. It is however very hard to obtain a proof because of the growing dimensionality of the underlying state space and the problem set down by the departure term. This proof is related to a well known fundamental queueing control problem, for which no rigorous proof does exist yet. We believe that our proof for the two-server case should give some indications that would motivate future research. This open question consists in showing the propagation of a monotonicity relation through the minimizing operator. In Remark 1 inside the proof of Theorem 1 in Appendix A, we provide the mathematical details of what should be proven to rigorously obtain the multi-server result. It reduces to that for the well known routing problem in the heterogeneous multi-server queue, where the objective is to find a non-preemptive routing policy that minimizes the long run average time in the system (Hajek, 1984; Lin and Kumar, 1984; de Véricourt and Zhou, 2005). For a background on this question, we refer the reader to Koole (2007).

## 3.1 Preliminary Result

Proposition 1 provides a preliminary result for Model G.

**Proposition 1** *In the multi-server case of Model G, if all agents are idling and queue 2 is not empty, then it is optimal to serve at least an outbound call.*

***Proof.*** For $\gamma_2 > 0$, it is clear that an outbound call in queue 2 has to be served at one point. Otherwise, queue 2 would contain an infinite number of outbounds due to the FCFS rule. Therefore, a policy which would not serve an outbound call can not be optimal. We next prove that the best situation for the service of an outbound call is when all agents are idling. Serving an outbound call always improves the performance of outbounds whether this outbound call is served when all agents are idling or in another situation. An outbound taken in service would deteriorate the performance of inbounds if new inbounds arrive at a busy system while this outbound call is still in service. The lowest value of the probability of such an event is reached in the case this outbound call has been taken in service when all agents are idling. Moreover, an outbound call service duration does not depend on the system state. Thus, serving an outbound call when all agents are idling improves the performance of outbounds and has the smallest probability to deteriorate the performance measures of inbounds. Since all outbounds has to be served at one point, an optimal state-dependent policy forces the service of outbounds, if any, when all agents are idle. □

We next deduce the optimal agent reservation policy for the single-server case of Model G.

**Corollary 1** *In the single-server case of Model G, the optimal agent reservation policy is the non-idling policy.*

The proof of Corollary 1 directly follows from Proposition 1. In Section 1 of the online supplement, we propose another proof of this corollary for Model A using an MDP approach.

## 3.2 Two-server Result for Model A

In the two-server case, using Proposition 1, we never encounter situations for the optimal policy where the two agents are idling and at least one outbound call is in queue 2. When one server is busy, we prove in Theorem 1 that the optimal policy in Model A is of threshold type for the reservation of the other server.

**Theorem 1** *In the two-server case for Model A, when one agent is busy, there exists a threshold on the number of outbounds in queue 2, at and beyond which it is optimal to serve the first waiting outbound in line, and it is optimal to not serve outbounds in the remaining cases.*

The proof is given in Appendix A. It is based on the propagation of monotonicity results of the value function as defined in Section 2. This type of proofs is standard in MDP problems (Koole, 2007). Yet, our result can not directly follow from Koole (2007) for the following reasons. The existing results concern mostly the single-server-one-dimensional case. Less is doable in the multi-dimensional case for the propagation of the results through the minimizing operator. Moreover, abandonment from queue 1 is allowed here, a feature that often breaks the monotonicity properties when space truncation is required. We show in our proof that the monotonicty properties are maintained. Finally, the complexity of the proof comes from the arrival term, which is specific in our model and requires a special consideration, because the two queues are involved and the customer reaction is state-dependent.

## 3.3 Multi-Server Conjecture

Let us now comeback to the multi-server case. Using the value functions defined in Section 2, we conjecture that the optimal policy is of switch type. For both Models A and G, we conduct a numerical study from which we deduce the switching curves which separate states where it is optimal to serve an outbound call from those where it is not. We also examine the impact of the system parameters on the reservation policy.

### 3.3.1 Switching Curves for Model A

For Model A, we do not need to distinguish between inbounds and outbounds in service. Let us rewrite the value functions for Model A ($\mu_1 = \mu_2 = \mu$, $r_1 = 1$). We have for $n \geq 0$,

$$\begin{aligned}
U_{n+1}(x,y) = \gamma_1(x-s)^+ + \gamma_2 y + \lambda \big[&\mathbf{1}_{(0\leq x<s)} V_n(x+1,y) \ +\mathbf{1}_{(s\leq x<s+k)} \left((1-\alpha_x)V_n(x+1,y) + \alpha_x(V_n(x,y)+\gamma_3)\right) \\
&+ \mathbf{1}_{(s+k\leq x<s+N)}(q_x V_n(x,y+1) + \alpha_x(V_n(x,y)+\gamma_3) + (1-q_x-\alpha_x)V_n(x+1,y)) \\
&+\mathbf{1}_{(x=s+N)}(q_{N-1}V_n(x,y+1) + (1-q_{N-1})(V_n(x,y)+\gamma_3))\big] \\
+ \beta(x-s)^+(V_n(x-1,y)+\gamma_3) &+ \min(s,x)\mu V_n(x-1,y) + \big(1-\lambda-\beta(x-s)^+ - \min(s,x)\mu_1\big) V_n(x,y), \text{ for } x,y \geq 0,
\end{aligned}$$

with

$$V_{n+1}(x,y) = \min(U_{n+1}(x+1,y-1), U_{n+1}(x,y)),$$

for $y > 0$ and $0 \leq x < s$ and $V_{n+1}(x,y) = U_{n+1}(x,y)$ in the remaining cases. We choose $V_0(x,y) = U_0(x,y) = 0$, for $x, y \geq 0$.

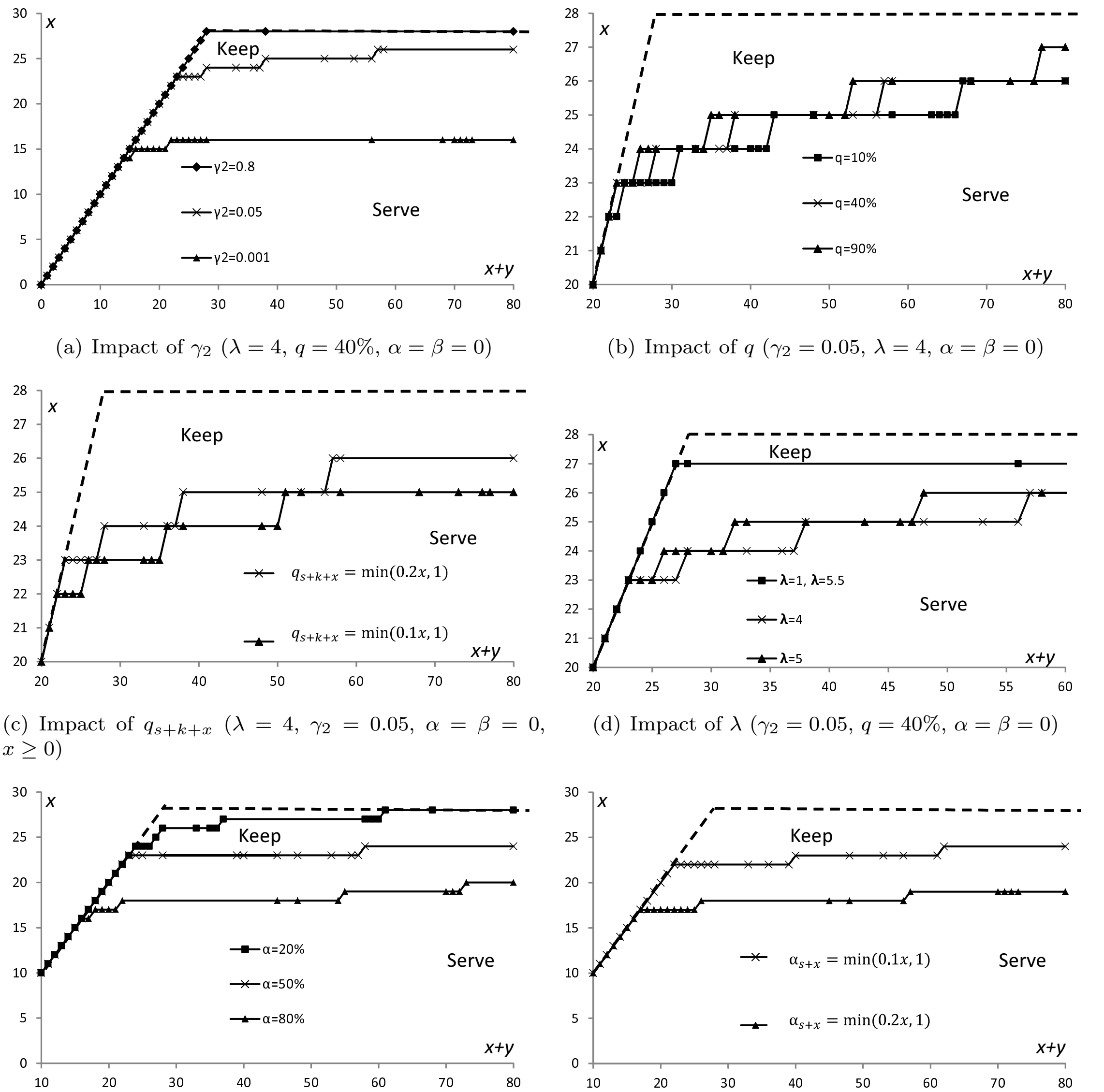
We conjecture that the optimal policy is a function of $x$ (number of calls in service plus number of inbounds in queue 1) and $y$ (number of outbounds in queue 2). Figure 4 gives various optimal switching curves to illustrate the impact of the system parameters on the optimal policy. The abscissa axis in each figure represents the overall number of jobs in the system (number of outbounds in queue 2 plus number of calls in service) and the ordinate axis represents the number of calls in service. We only consider states where $0 \leq x < s$. For the remaining states, the only possible action is to keep outbounds in the queue. The optimal actions can be read from the figures. Consider a given point $(x+y, x)$ $(0 \leq x < s$ and $y > 0)$. If this point is strictly under the curve, then it is optimal to serve an outbound call and therefore move from $(x+y, x)$ to $(x+1+y-1, x+1) = (x+y, x+1)$. If this new point is strictly under the curve then the optimal action is to serve another outbound call. We continue to take the decision *to serve* by moving on a vertical line until we reach the curve. On the switching curve or above, the optimal action is to keep outbounds in the queue. The value to choose $x+y$ in abscissa instead of $y$ is to observe the evolution from a non-optimal point to the optimal one on a vertical line instead of a diagonal one. The curves in dashed lines represent the non-idling policy.

We observe that when $x = 0$ and $y > 0$, the optimal action is always to serve an outbound call (this holds from Proposition 1). Given that the switching curve is increasing in $x+y$, it is an increasing step function. It is given by

$$c(x+y) = \min(y_0, x+y) + \mathbf{1}_{(x+y \geq y_1)} + \mathbf{1}_{(x+y \geq y_2)} + \cdots + \mathbf{1}_{(x+y \geq y_{s-y_0})}, \tag{2}$$

where $1 \leq y_0 < y_1 < y_2 < \cdots < y_{s-y_0}$. The parameters $y_0$, ..., $y_{s-y_0}$ are the levels that represent the changing points of the switching curve. Using Proposition 1, we have $y_0 \neq 0$. Equation (2) can be interpreted as follows. Assume we have $x+y$ jobs in the system ($x$ busy agents and $y$ outbounds in queue 2). If $x+y < y_1$, then it is optimal to have at most $y_0$ tasks in service, i.e., if $x < y_0$ we move from state $(x,y)$ to state $(\min(y_0, x+y), y - (\min(y_0, x+y) - x))$, and if $x \geq y_0$ we stay in state $(x,y)$. If $y_1 \leq x+y < y_2$, then at most $y_0+1$ jobs should be in service, i.e., if $x < y_0+1$ we move from state $(x,y)$ to state $(\min(y_0+1, x+y), y - (\min(y_0+1, x+y) - x))$, and if $x \geq y_0+1$ we stay in state $(x,y)$, and so on. Finally, if $y \geq y_{s-y_0}$, then at most $y_0 + s - y_0 = s$ jobs should be in service. In other words, when $x+y \geq y_{s-y_0}$, no agents are reserved for inbounds and it is optimal to move from state $(x,y)$ to state $(\min(s, x+y), y - (\min(s, x+y) - x))$. A qualitative interpretation of Equation (2) is that the more numerous queued outbounds and the less busy are the agents, the more likely the optimal decision would be to serve an outbound call.

This switch type policy in the multi-server case is a standard extension of the threshold policy in the two-server case. The new element in the multi-server case is that the decision to serve an outbound call should no longer only depend on the length of queue 2, since more than one agent might be involved. For



(a) Impact of $\gamma_2$ ($\lambda = 4$, $q = 40\%$, $\alpha = \beta = 0$)

(b) Impact of $q$ ($\gamma_2 = 0.05$, $\lambda = 4$, $\alpha = \beta = 0$)

(c) Impact of $q_{s+k+x}$ ($\lambda = 4$, $\gamma_2 = 0.05$, $\alpha = \beta = 0$, $x \geq 0$)

(d) Impact of $\lambda$ ($\gamma_2 = 0.05$, $q = 40\%$, $\alpha = \beta = 0$)

(e) Impact of $\alpha$ ($\lambda = 4$, $\gamma_2 = 0.05$, $q = 40\%$, $\beta = 0.1$)

(f) Impact of $\alpha_{s+x}$ ($\lambda = 4$, $\gamma_2 = 0.05$, $q = 40\%$, $\beta = 0$, $x \geq 0$)

Figure 4: Optimal switching curve ($\mu_1 = 0.2$, $r_1 = 1$, $s = 28$, $\gamma_1 = 1$, $k = 5$, $\gamma_3 = 0.5$)

a given situation with $x$ busy agents and $s - x$ idle agents, the optimal policy is a threshold policy on the length of queue 2. This leads, as a consequence, to a switch type policy.

We next examine the impact of the parameters on the reservation policy. In Proposition 2, we prove that the more importance is given to inbounds and the less customers are likely to accept the callback offer, the higher should be the reservation for inbounds.

**Proposition 2** *Consider two situations with identical arrival and departure parameters ($\lambda$, $\alpha_x$ for $x \geq s$, $\beta$, $s$ and $\mu$). The first situation has the cost parameters $\gamma_1$, $\gamma_2$ and $\gamma_3$ and the second one has $\gamma_1'$, $\gamma_2'$ and $\gamma_3'$. The callback parameters are constant for both situations. They are $q$ and $q + q'$ for the first and second situations, respectively.*

*If $\gamma_1 \geq \gamma_1'$, $\gamma_2 \leq \gamma_2'$, $\gamma_3 \geq \gamma_3'$, $q' \geq 0$, then the first situation requires more reservation than the second one. In other words, the switching curve is lower for the first situation.*

The proof of this proposition is given in Appendix B. The impact of the cost parameter $\gamma_2$ is illustrated in Figure 4(a), i.e., the switching curve increases (the reservation decreases) in $\gamma_2$. The opposite is true when $\gamma_1$ or $\gamma_3$ increases. Figure 4(b) illustrates the impact of a constant callback parameter ($q_x = q$ for $x \geq s+k$). It shows that the more customers are likely to accept the callback option, the higher is the switching curve (less reservation for inbounds). The same observation holds when $q_x$ is not constant (Figure 4(c)). The key factor, wether the callback parameter is constant or not, is the proportion of outbounds.

A less intuitive observation is that the switching curve is not monotone in the workload, defined as $\lambda/\mu$ (Figure 4(d)). We observe that reservation does not happen in the extreme situations of light or heavy workloads. For light workload situations, the system capacity is high enough, such that both call types experience small waiting times. Then, the reservation for inbound calls does not need to be substantial. For high workload situations, queue 1 is often long. Thus, a high proportion of customers would choose the callback option and join queue 2. Given that queue 2 is also long, the system should not further deteriorate the waiting of outbounds by reserving agents for jobs 1. However, for an intermediate situation, with a moderate workload, jobs 2 are less numerous, and do not therefore need to have access to all agents. The system may then consider agent reservation for jobs 1.

Figure 4(e) reveals that the impact of the balking parameters $\alpha_x$ and the abandonment parameter $\beta$ are not similar to that of the workload. For high values of $\alpha_x$ or $\beta$, the system capacity is high enough to achieve small waiting times. However, the proportion of abandonment is high, so, the reservation for inbounds needs to be important to avoid too much abandonment. For low values of $\alpha_x$ or $\beta$, the reservation policy mainly depends on the workload $\lambda/\mu$ (see Figures 4(e) and 4(f)).

### 3.3.2 Switching Curves for Model G

We now consider Model G. Figures 5 and 6 illustrate the switching curves for the optimal policy in Model G. Again, the curves in dashed lines represent the non-idling policy.

As expected, we observe that the optimal decisions are not only based on the number of outbounds in queue 2 and the number of busy agents as for Model A, but also the identity of the jobs in service. We distinguish three different zones delimited by two switching curves. A first switching curve is defined for the case where all busy agents are busy with rate $\mu_1$ ($s_2 = s_3 = 0$). This situation is the worst for the occupancy of the agents, because $\mu_3 \geq \mu_2 \geq \mu_1$. Thus, under this first switching curve, for any state with less busy agents or more outbounds in queue 2, the optimal decision is to serve an outbound call (if any), i.e., we move from state $(x+y+s_2+s_3, x+s_2+s_3)$ to state $(x+1+y-1+s_2+s_3, x+1+s_2+s_3) = (x+y+s_2+s_3, x+1+s_2+s_3)$. A second switching curve is defined for the cases were all busy agents are busy with rate $\mu_3$ ($x = s_2 = 0$). This situation is the best for the occupancy of the agents. On and above this second switching curve, for any state with more busy agents and less outbounds in queue 2, the optimal decision is to keep all outbounds in queue 2.

The ordering $\mu_3 \geq \mu_2 \geq \mu_1$ justifies that the first switching curve is below the second one. Even in the

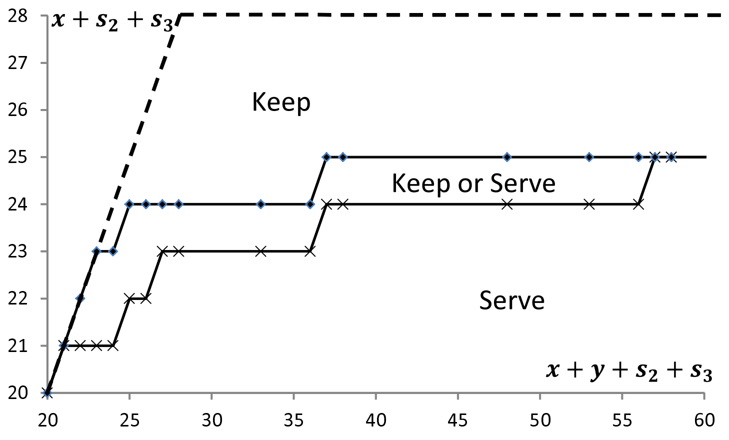
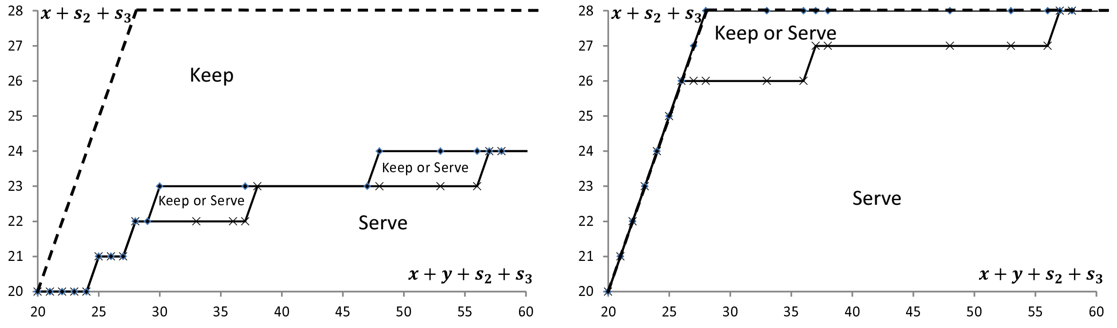


Figure 5: Optimal switching curves ($\lambda = 3.8$, $q = 40\%$, $\alpha = \beta = 0$, $\gamma_1 = 1$, $\gamma_2 = 0.05$, $k = 5$, $\mu_1 = 0.2$, $\mu_2 = 1$, $\mu_3 = 10$, $r_1 = r_2 = 1/3$, $s = 28$)

case $\mu_3 = \mu_2$, the second switching curve ($x = s_2 = 0$) is still higher than a switching curve where all busy agents are busy with rate $\mu_2$ ($x = s_3 = 0$). The reason is the high need of serving outbounds when all agents are busy with rate $\mu_3$. If the agents are all handling a non-available outbound situation, they would not reduce the number of outbounds in the system, so, the need for serving outbounds does not reduce.

Yet, for situations with small number of customers in the system or high number of customers in queue 2, the two extreme switching curves (corresponding to $s_2 = s_3 = 0$ and $x = s_2 = 0$) coincide. Therefore, there only exists a finite number of states where the optimal decisions depend on the identity of the jobs in service. Figure 6(a) reveals that the two extreme switching curves get closer to one another as $r_1$, $\mu_2$, or $\mu_3$ increases. The reason is the similarity between the service requirements of inbounds and outbounds. Figure 6(b) reveals that as $r_1 + r_2$ decreases, the two extreme switching curves get higher, i.e., less agent reservation. The reason is related to the difficulty of serving an outbound call. When agents are often handling non-available outbound situation, it is difficult to reduce the length of queue 2, therefore, outbounds should benefit from more availability of the agents.



(a) Example with $r_1 = 80\%$, $r_2 = 5\%$, $\mu_1 = 0.2$, $\mu_2 = 0.5$ and $\mu_3 = 10$

(b) Example with $r_1 = 10\%$, $r_2 = 10\%$, $\mu_1 = 0.2$, $\mu_2 = 1$ and $\mu_3 = 10$

Figure 6: Optimal switching curve ($\lambda = 3.8$, $q = 40\%$, $\alpha = \beta = 0$, $\gamma_1 = 1$, $\gamma_2 = 0.05$, $k = 5$, $s = 28$)

Similarly to Model A, since the switching curve is increasing in $x + y + s_2 + s_3$, it is an increasing step function. Given that agents handle 3 different types of jobs, we define the 3 variables increasing function $b(x, s_2, s_3)$ which gives the "busyness" of the agents team. Because the number of agents is finite, we assume without loss of generality that $0 \leq b(x, s_2, s_3) \leq 1$. This busyness function corrects the switching curve, defined for Model A, into

$$c(x + y + s_2 + s_3) = \min(y_0, x + y + s_2 + s_3)\mathbf{1}_{(b(x,s_2,s_3)\leq b_0)} + \mathbf{1}_{(x+y+s_2+s_3\geq y_1)}\mathbf{1}_{(b(x,s_2,s_3)\leq b_1)}$$
$$+ \mathbf{1}_{(x+y+s_2+s_3\geq y_2)}\mathbf{1}_{(b(x,s_2,s_3)\leq b_2)} + \cdots + \mathbf{1}_{(x+y+s_2+s_3\geq y_{s-y_0-1})}\mathbf{1}_{(b(x,s_2,s_3)\leq b_{s-y_0-1})} + \mathbf{1}_{(x+y+s_2+s_3\geq y_{s-y_0})},$$

where $1 \leq y_0 < y_1 < y_2 < \cdots < y_{s-y_0}$ and $0 < b_0 \leq b_1 \leq \cdots \leq b_{s-y_0-1} \leq 1$. The parameters $y_i$, $0 \leq i \leq s - y_0$, have the same signification as those for Model A. The parameters $b_i$, $0 \leq i \leq s - y_0$, are the levels of change of the busyness of the agents team. The values of the $b_i$s can be determined using value iteration.

From the numerical experiments, we observe that the values of the $b_i$s are different than one only for small values of $i$. This implies that the busyness of the agents team affects the optimal decisions only when the number of busy agents is low. The reason is related to the blocking risk for an inbound call. When most of the agents are idling, the decision to serve an outbound call would most likely not block the agents team. In such a situation, what affects the decision is then the identity of jobs in service. In the opposite case, when most of the agents are busy, the service of an outbound call could easily lead to a blocking situation (waiting time for inbound calls). In such a situation, what affects the decision is then the total number of busy agents $(x + s_2 + s_3)$ and the length of queue 2 $(y)$, more than the identity of the jobs in service.

# 4 Performance Analysis

We compute the stationary performance measures. In Section 4.1, we profit from the constant transition rates and propose an exact algorithm for Model C. In Section 4.2, we provide a controlled approximation based on value iterations for Models A and G. In Section 4.3, we consider special cases of agent reservation for Model C (Section 4.3.1) and the non-idling case for Model A (Section 4.3.2). This allows to obtain closed-form expressions for the bounds of the performance measures of Models A and C.

## 4.1 Model C

We compute here $E(W_1)$, $E(W_2)$, $P_b$ and $\Psi$. Our approach is based on the analysis of the underlying Markov chain. We compute the stationary probabilities of the system states by solving a system of linear difference equations. We do so by solving the involved homogeneous equations defined on the set of complex numbers. Although some quantities contain infinite summations, we provide a method that allows to do the exact computation within a finite number of calculations.

Consider the stochastic process $\{(x(t), y(t)), t \geq 0\}$, where $x(t)$ denotes the number of calls in queue 1 (jobs 1) or in service (jobs 1 or 2); and $y(t)$ denotes that in queue 2 (jobs 2) at a given time $t \geq 0$. We have $x(t), y(t) \in \{0, 1, 2, ...\}$, for $t \geq 0$. As inter-arrival and service times are exponentially distributed, $\{(x(t), y(t)), t \geq 0\}$ is a Markov chain. An illustration of this Markov chain in given in Figure 7.

We denote by $p_{x,y}$ the stationary probability to be in state $(x, y)$, for $x, y \in \mathbb{N}$. In what follows, we

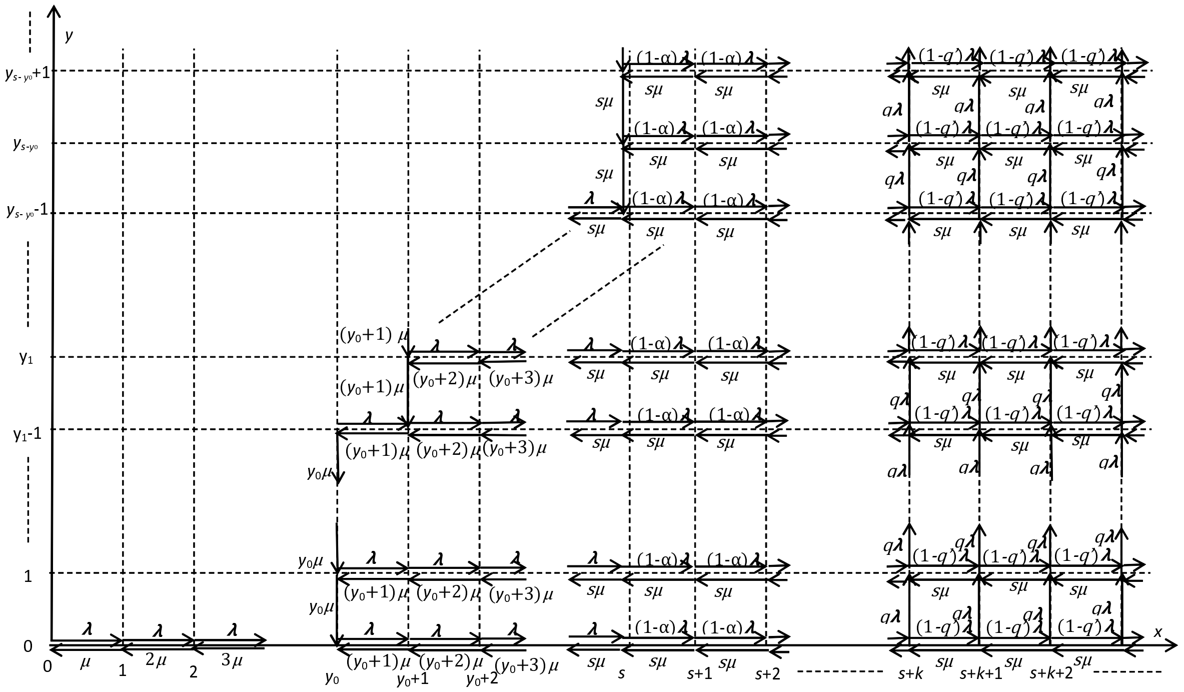


Figure 7: Markov chain for Model C ($q' = q + \alpha$)

compute the stationary probabilities, from which we thereafter deduce the system performance measures of interest. To simplify the presentation of the analysis, we divide it into the following 7 steps:

- **Step 1.** We provide the set of equilibrium equations relating the stationary probabilities.
- **Step 2.** We simplify the expressions of $p_{x,y}$, for $x \leq s+k$ and $y \geq 0$, by expressing them as a function of only two state probabilities from the row $y$ in the Markov chain.
- **Step 3.** We show how $p_{x,y}$, for $x \geq s+k$ and $y \geq 0$, can be computed as a function of $p_{s+k,0}, p_{s+k,1}, \cdots, p_{s+k,y}$.
- **Step 4.** We evaluate all stationary probabilities for $x \geq 0$ and $y = 0$ as a function of $p_{0,0}$.
- **Step 5.** For $y \geq 0$, we develop a recurrence method to compute all stationary probabilities of row $y+1$ in the Markov chain as a function of the previous rows. Thus all stationary probabilities can be derived as a function of $p_{0,0}$.
- **Step 6.** Although $p_{0,0}$ involves an infinite summation, we provide a method to compute it within a finite number of calculations.
- **Step 7.** We finally derive the system performance measures as a function of the stationary probabilities.

The details for each step are given in Appendix C.

## 4.2 Models A and G

We compute here $E(W_1)$, $E(W_2)$, $P_a$, and $\Psi$. We propose a numerical method based on the iterative computation of the dynamic programming operators.

For Model A, assuming the switch policy as defined in Section 3.3.1, the value functions can be rewritten, for $n \geq 0$, as

$$\begin{aligned}
V_{n+1}(x,y) = {} & \gamma_1(x-s)^+ + \gamma_2 y + \lambda \big[\mathbf{1}_{(0\leq x<s)} V_n(x+1,y) \ +\mathbf{1}_{(s\leq x<s+k)} \left((1-\alpha_x)V_n(x+1,y) + \alpha_x(V_n(x,y)+\gamma_3)\right) \\
& + \mathbf{1}_{(s+k\leq x<s+N)}(q_x(V_n(x,y+1)+\gamma_4) + \alpha_x(V_n(x,y)+\gamma_3) + (1-q_x-\alpha_x)V_n(x+1,y)) \\
& +\mathbf{1}_{(x=s+N)}(q_{N-1}(V_n(x,y+1)+\gamma_4) + (1-q_{N-1})(V_n(x,y)+\gamma_3))\big] \\
& + \beta(x-s)^+(V_n(x-1,y)+\gamma_3) \\
& + \min(s,x)\mu \Big[\mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y\leq y_1, x\leq y_0)} + \mathbf{1}_{(y_1<x+y\leq y_2, x\leq y_0+1)} + \mathbf{1}_{(y_2<x+y\leq y_3, x\leq y_0+2)} + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y, x\leq s)})V_n(x,y-1) \\
& + \Big(1 - \mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y\leq y_1, x\leq y_0)} + \mathbf{1}_{(y_1<x+y\leq y_2, x\leq y_0+1)} + \mathbf{1}_{(y_2<x+y\leq y_3, x\leq y_0+2)} + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y, x\leq s)})\Big) V_n(x-1,y)\Big] \\
& + \left(1-\lambda-\beta(x-s)^+ - \min(s,x)\mu_1\right) V_n(x,y), \ \text{for } x,y\geq 0,
\end{aligned}$$

with $V_0(x,y)=0$, for $x,y\geq 0$.

For Model G, assuming the switch policy as defined in Section 3.3.2, the value functions can be rewritten, for $n \geq 0$, as

$$\begin{aligned}
& U_{n+1}(x,y,s_2,s_3) = \gamma_1(x+s_2+s_3-s)^+ + \gamma_2 y \\
& + \lambda \big[\mathbf{1}_{(0\leq x+s_2+s_3<s)} V_n(x+1,y,s_2,s_3) \\
& \qquad +\mathbf{1}_{(s\leq x+s_2+s_3<s+k)} \left((1-\alpha_{x,s_2,s_3})V_n(x+1,y,s_2,s_3) + \alpha_{x,s_2,s_3}(V_n(x,y,s_2,s_3)+\gamma_3)\right) \\
& \qquad + \mathbf{1}_{(s+k\leq x+s_2+s_3<s+N)}(q_{x,s_2,s_3}(V_n(x,y+1,s_2,s_3)+\gamma_4) + \alpha_{x,s_2,s_3}(V_n(x,y,s_2,s_3)+\gamma_3) \\
& \qquad + (1-q_{x,s_2,s_3}-\alpha_{x,s_2,s_3})V_n(x+1,y,s_2,s_3)) \\
& \qquad +\mathbf{1}_{(x+s_2+s_3=s+N)}(q_{N-1,s_2,s_3}(V_n(x,y+1,s_2,s_3)+\gamma_4) + (1-q_{N-1,s_2,s_3})(V_n(x,y,s_2,s_3)+\gamma_3))\big] \\
& + \beta(x+s_2+s_3-s)^+(V_n(x-1,y,s_2,s_3)+\gamma_3) \\
& + \min(s-s_2-s_3,x)\mu_1 \big[\mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y+s_2+s_3\leq y_1, x+s_2+s_3\leq y_0, b(x-1,s_2,s_3)\leq b_0)} + \mathbf{1}_{(y_1<x+y+s_2+s_3\leq y_2, x+s_2+s_3\leq y_0+1, b(x-1,s_2,s_3)\leq b_1)} \\
& \qquad + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y+s_2+s_3, x+s_2+s_3\leq s)})(r_1 V_n(x,y-1,s_2,s_3) + r_2 V_n(x-1,y-1,s_2+1,s_3) \\
& \qquad + (1-r_1-r_2)V_n(x-1,y-1,s_2,s_3+1)) \\
& \qquad + \big(1 - \mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y+s_2+s_3\leq y_1, x+s_2+s_3\leq y_0, b(x-1,s_2,s_3)\leq b_0)} + \mathbf{1}_{(y_1<x+y+s_2+s_3\leq y_2, x+s_2+s_3\leq y_0+1, b(x-1,s_2,s_3)\leq b_1)} \\
& \qquad + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y+s_2+s_3, x+s_2+s_3\leq s)})\big) V_n(x-1,y,s_2,s_3)\big] \\
& + s_2\mu_2 \big[\mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y+s_2+s_3\leq y_1, x+s_2+s_3\leq y_0, b(x,s_2-1,s_3)\leq b_0)} + \mathbf{1}_{(y_1<x+y+s_2+s_3\leq y_2, x+s_2+s_3\leq y_0+1, b(x,s_2-1,s_3)\leq b_1)} \\
& \qquad + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y+s_2+s_3, x+s_2+s_3\leq s)})(r_1 V_n(x+1,y-1,s_2-1,s_3) + r_2 V_n(x,y-1,s_2,s_3) \\
& \qquad + (1-r_1-r_2)V_n(x,y-1,s_2-1,s_3+1)) \\
& \qquad + \big(1 - \mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y+s_2+s_3\leq y_1, x+s_2+s_3\leq y_0, b(x,s_2-1,s_3)\leq b_0)} + \mathbf{1}_{(y_1<x+y+s_2+s_3\leq y_2, x+s_2+s_3\leq y_0+1, b(x,s_2-1,s_3)\leq b_1)} \\
& \qquad + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y+s_2+s_3, x+s_2+s_3\leq s)})\big) V_n(x,y,s_2-1,s_3)\big] \\
& + s_3\mu_3 \big[\mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y+s_2+s_3\leq y_1, x+s_2+s_3\leq y_0, b(x,s_2,s_3-1)\leq b_0)} + \mathbf{1}_{(y_1<x+y+s_2+s_3\leq y_2, x+s_2+s_3\leq y_0+1, b(x,s_2,s_3-1)\leq b_1)} \\
& \qquad + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y+s_2+s_3, x+s_2+s_3\leq s)})(r_1 V_n(x+1,y-1,s_2,s_3-1) + r_2 V_n(x,y-1,s_2+1,s_3-1) \\
& \qquad + (1-r_1-r_2)V_n(x,y-1,s_2,s_3)) \\
& \qquad + \big(1 - \mathbf{1}_{(y>0)}(\mathbf{1}_{(x+y+s_2+s_3\leq y_1, x+s_2+s_3\leq y_0, b(x,s_2,s_3-1)\leq b_0)} + \mathbf{1}_{(y_1<x+y+s_2+s_3\leq y_2, x+s_2+s_3\leq y_0+1, b(x,s_2,s_3-1)\leq b_1)} \\
& \qquad + \cdots + \mathbf{1}_{(y_{s-y_0}<x+y+s_2+s_3, x+s_2+s_3\leq s)})\big) V_n(x,y,s_2,s_3-1)\big] \\
& + \left(1-\lambda-\beta(x+s_2+s_3-s)^+ - \min(s-s_2-s_3,x)\mu_1 - s_2\mu_2 - s_3\mu_3\right) V_n(x,y,s_2,s_3), \ \text{for } x,y\geq 0, \ \text{and } 0\leq s_2,s_3\leq s,
\end{aligned}$$

with $V_0(x,y,s_2,s_3)=0$, for $x,y\geq 0$ and $0\leq s_2,s_3\leq s$.

In both cases (Models A and G), the standard way of obtaining the long-term performance measures is through value iteration, by recursively evaluating $V_n$, for $n\geq 0$. As $n$ tends to infinity, the difference $V_{n+1}(x,y,s_2,s_3)-V_n(x,y,s_2,s_3)$ converges to the desired metric. Thus, we stop the iteration until the following criterion is met

$$\max_{x,y,s_2,s_3}\{V_{n+1}(x,y,s_2,s_3)-V_n(x,y,s_2,s_3)\}-\min_{x,y,s_2,s_3}\{V_{n+1}(x,y,s_2,s_3)-V_n(x,y,s_2,s_3)\}<\epsilon,$$

for some given small $\epsilon$.

In what follows we precise the parameters in the value functions which allow to compute the desired performance measures. One can calculate the expected number of customers in queue 1, say $E(N_1)$, by letting $\gamma_1=1,\gamma_2=0,\gamma_3=0,\gamma_4=0$ in the value function; the expected number of customers in queue 2, say $E(N_2)$, by letting $\gamma_1=0,\gamma_2=1,\gamma_3=0,\gamma_4=0$; the proportion of customers who abandon the system, $P_a$, by letting $\gamma_1=0,\gamma_2=0,\gamma_3=1/\lambda,\gamma_4=0$; the proportion of customers who choose the callback offer, $\Psi$, by letting $\gamma_1=0,\gamma_2=0,\gamma_3=0,\gamma_4=1/\lambda$. Using next the Little law, we obtain the expected waiting time for served customers in queue 1, $E(W_1)=\frac{E(N_1)}{\lambda(1-P_a-\Psi)}$; and the expected waiting time in queue 2, $E(W_2)=\frac{E(N_2)}{\lambda\Psi}$.

## 4.3 Special Cases

We consider here some special cases of agent reservation for Model C and the non-idling case for Model A.

### 4.3.1 Special Reservation Cases for Model C

We define for Model C the threshold $y_0$ on the number of busy agents. If the number of busy agents is lower than or equal to $y_0$ $(1\leq y_0\leq s)$ and at least one outbound call is in queue 2, then we serve this outbound call. In the remaining cases, we do not serve outbounds. Therefore, the switching curve of this policy is $c(x+y)=\min(x+y,y_0)$.

Since the optimal action is to serve an outbound call when all agents are idling (Proposition 1), the worst policy for outbounds (the best case for inbounds) consists of serving an outbound call only when all agents are idling. We refer to the latter as the highest reservation policy. It corresponds to the case $y_0=1$. As for the non-idling policy, it corresponds to the case $y_0=s$.

The analysis of this policy is a deduced from that of Section 4.1. In Corollary 2, we give closed-form expressions for $E(W_1)$, $P_b$ and $\Psi$ as a function of $y_0$. The proof is given in Section 2 of the online supplement.

**Corollary 2** *For $1\leq y_0\leq s$, we have*

$$\Psi=\frac{q\left(\frac{a(1-\alpha)}{s}\right)^k\frac{a^s}{s!}}{1-\frac{a(1-q-\alpha)}{s}}\frac{p_{0,0}}{1-q\frac{a}{y_0}\frac{a^{s-y_0}y_0!}{s!}\left(\frac{a(1-\alpha)}{s}\right)^k\frac{1}{1-\frac{a(1-q-\alpha)}{s}}},$$

$$P_b = \frac{\alpha \frac{a^s}{s!} p_{0,0} \left( \frac{1-\left(\frac{a(1-\alpha)}{s}\right)^k}{1-\frac{a(1-\alpha)}{s}} + \frac{\left(\frac{a(1-\alpha)}{s}\right)^k}{1-\frac{a(1-q-\alpha)}{s}} \right)}{1 - q\frac{a}{y_0} \frac{a^{s-y_0} y_0!}{s!} \left(\frac{a(1-\alpha)}{s}\right)^k \frac{1}{1-\frac{a(1-q-\alpha)}{s}}},$$

$$E(W_1) = \frac{\frac{a^s}{s!}}{\lambda(1-\Psi-P_b)} \frac{p_{0,0} \left( \sum_{x=0}^{k-1} x \left(\frac{a(1-\alpha)}{s}\right)^x + \left(\frac{a(1-\alpha)}{s}\right)^k \left( \frac{k\left(1-\frac{a(1-q-\alpha)}{s}\right) + \frac{a(1-q-\alpha)}{s}}{\left(1-\frac{a(1-q-\alpha)}{s}\right)^2} \right) \right)}{1 - q\frac{a}{y_0} \frac{a^{s-y_0} y_0!}{s!} \left(\frac{a(1-\alpha)}{s}\right)^k \frac{1}{1-\frac{a(1-q-\alpha)}{s}}},$$

*with*

$$p_{0,0} = \left[ \sum_{x=0}^{y_0-1} \frac{a^x}{x!} + \frac{\left( \sum_{x=0}^{s-y_0-1} \frac{a^{x+y_0}}{(y_0+x)!} + \frac{a^s}{s!} \sum_{x=0}^{k-1} \frac{(a(1-\alpha))^x}{s^x} + \frac{a^s}{s!} \frac{(a(1-\alpha))^k}{s^k} \frac{1}{1-\frac{a(1-q-\alpha)}{s}} \right)}{1 - q\frac{a}{y_0} \frac{a^{s-y_0} y_0!}{s!} \frac{(a(1-\alpha))^k}{s^k} \frac{1}{1-\frac{a(1-q-\alpha)}{s}}} \right]^{-1}.$$

In Appendix D, further simplifications of the above expressions are given for the multi-server special cases: $y_0 = 1$ (highest reservation) and $y_0 = s$ (non-idling).

### 4.3.2 Non-idling Case for Model A

We provide in Proposition 3 closed-form expressions for $E(W_1)$, $P_b$, $P_a$ and $\Psi$. The proof is given in Section 3 of the online supplement.

**Proposition 3** *For the non-idling case, we have*

$$p_{0,0} = \left[ \sum_{x=0}^{s-1} \frac{a^x}{x!} + \frac{\frac{a^s}{s!} \left( \sum_{x=0}^{k} \lambda^x \prod_{i=1}^{x} \left(\frac{1-\alpha_{i-1}}{s\mu+i\beta}\right) + \sum_{x=1}^{\infty} \lambda^{x+k} \frac{\prod_{i=1}^{k}(1-\alpha_{i-1}) \prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)} \right)}{1 - \frac{a}{s}\prod_{i=1}^{k}(1-\alpha_{i-1}) \sum_{x=0}^{\infty} \frac{q_{k+x}\lambda^{x+k} \prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)}} \right]^{-1},$$

$$\Psi = \sum_{x=0}^{\infty} q_{k+x} P_{s+k+x} = \frac{s}{a}(P_s - p_{s,0}) = \frac{\frac{a^s}{s!}\prod_{i=1}^{k}(1-\alpha_{i-1}) \sum_{x=0}^{\infty} \frac{q_{k+x}\lambda^{x+k} \prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)}}{1 - \frac{a}{s}\prod_{i=1}^{k}(1-\alpha_{i-1}) \sum_{x=0}^{\infty} \frac{q_{k+x}\lambda^{x+k} \prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)}} p_{0,0},$$

$$P_b = \sum_{x=0}^{\infty} \alpha_x P_{s+x} = \frac{\frac{a^s}{s!} \left( \sum_{x=0}^{k} \alpha_x \lambda^x \prod_{i=1}^{x} \left(\frac{1-\alpha_{i-1}}{s\mu+i\beta}\right) + \sum_{x=1}^{\infty} \alpha_{k+x} \prod_{i=1}^{k}(1-\alpha_{i-1}) \frac{q_{k+x}\lambda^{x+k} \prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)} \right)}{1 - \frac{a}{s}\prod_{i=1}^{k}(1-\alpha_{i-1}) \sum_{x=0}^{\infty} \frac{q_{k+x}\lambda^{x+k} \prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)}} p_{0,0},$$

$$P_a = \frac{\frac{a^s}{s!}\left(\sum_{x=0}^{k}\left(\alpha_x + x\frac{\beta}{\lambda}\right)\lambda^x\prod_{i=1}^{x}\left(\frac{1-\alpha_{i-1}}{s\mu+i\beta}\right) + \sum_{x=1}^{\infty}\left(\alpha_{x+k} + (x+k)\frac{\beta}{\lambda}\right)\prod_{i=1}^{k}(1-\alpha_{i-1})\frac{q_{k+x}\lambda^{x+k}\prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)}\right)}{1-\frac{a}{s}\prod_{i=1}^{k}(1-\alpha_{i-1})\sum_{x=0}^{\infty}\frac{q_{k+x}\lambda^{x+k}\prod_{i=k+1}^{x+k}(1-\alpha_{i-1}-q_{i-1})}{\prod_{i=1}^{x+k}(s\mu+i\beta)}} p_{0,0},$$

*and*

$$E(W_1) = \frac{P_a - P_b}{\beta(1-\Psi-P_a)}.$$

# 5 Numerical Experiments

We investigate the benefits of the callback offer and the impact of the policy parameters on the system performance. The policy parameters are the state-dependent number of agents reserved for inbounds, and the threshold $k$ for the callback proposition. Because of the analysis complexity, the conclusions we derive are mainly based on numerical observations. For some particular cases, we develop analytical results that provide better understanding and support the conclusions.

## 5.1 Benefits of the Callback Offer

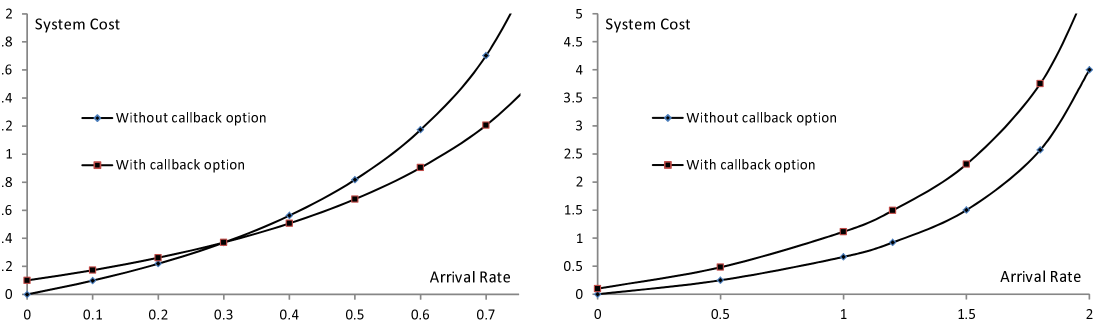
We evaluate the benefits of the callback offer on the performance measures, in relation with the system workload. We consider the single-server non-idling case (optimal policy, Corollary 1) of Model C with $k = 0$ (optimal $k$, Proposition 5). The objective is to provide a simple closed-form expression of the difference between the system cost in two situations; with and without the callback offer. Using Corollary 2, for $y_0 = s = 1$ and $k = 0$ (situation with the callback option), we obtain $P_b = \frac{\alpha a}{1+\alpha a}$, $\Psi = \frac{qa}{1+\alpha a}$, $E(W_1) = \frac{1}{\mu}\frac{(1-q-\alpha)a}{(1-(1-q-\alpha)a)(1-qa)}$, and $E(W_2) = \frac{1}{\mu}\frac{1+a(\alpha/q)(1-\alpha)(1-(1-q-\alpha)a)}{(1-a(1-\alpha))(1-(1-q-\alpha)a)}$. For $y_0 = s = 1$ and $k = \infty$ (situation without the callback option), we obtain after simplification $P_b = \frac{\alpha a}{1+\alpha a}$, $\Psi = 0$ and $E(W_1) = \frac{1}{\mu}\frac{(1-\alpha)a}{(1-(1-\alpha)a)}$. The last situation reduces to an M/M/1 queue with balking. We do not provide for it the expression of $E(W_2)$ because outbounds do not exist when the callback option is not offered (it is simply considered as zero). The difference in system costs between the two situations (without the offer minus with the offer), denoted by $\Delta(a)$, is then

$$\Delta(a) = \frac{\gamma_1}{\mu}\left(\frac{(1-\alpha)a}{(1-(1-\alpha)a)} - \frac{(1-q-\alpha)a}{(1-(1-q-\alpha)a)(1-qa)}\right) - \frac{\gamma_2}{\mu}\frac{1+a(\alpha/q)(1-\alpha)(1-(1-q-\alpha)a)}{(1-a(1-\alpha))(1-(1-q-\alpha)a)}.$$

This difference can be either positive or negative.

In Figure 8, we illustrate the impact of the arrival rate on the system cost in the two situations (with and without the callback offer). Figure 8(a) reveals that the callback option can improve the system cost when the arrival rate is high. Since we have $\Delta(0) = -\frac{\gamma_2}{\mu} < 0$, under light workload, the callback option should not be provided. Roughly speaking, if a call goes to queue 2, then this call would loose priority. This is not useful because calls in queue 1 have anyway short waiting times.

Under a heavy workload situation, the preference is not always for using the callback option. As $a$ tends to $\frac{1}{1-\alpha}$, $\Delta(a)$ becomes equivalent to $\frac{\gamma_1-\gamma_2\frac{1+\alpha a}{qa}}{\mu(1-(1-\alpha)a)}$. This expression can either be positive or negative depending on the sign of its numerator. The higher $\gamma_1$ or $q$ are in comparison with $\gamma_2$ and $\alpha$, the more likely this expression would be positive. More qualitatively, this implies that the callback offer has a positive effect only if a high importance is given to inbounds and if customers would easily accept the callback offer (Figure



(a) $\alpha = 10\%$ and $q = 40\%$

(b) $\alpha = 60\%$ and $q = 10\%$

Figure 8: System cost with and without the callback offer (non-idling case of Model C, $s = 1$, $k = 0$, $SC = E(W_1) + 0.1E(W_2)$)

8(a)). In the case where customers are more likely to balk than to accept the callback offer (Figure 8(b)), providing a callback offer would deteriorate the system cost.

Finally, we numerically observe that the function $|\Delta(a)|$ is increasing in $a$. This induces that the benefits or the loss due to the use of the callback option would be more apparent under a heavy workload situation. This is precisely the value of the callback offer, that could better manage congested situations.

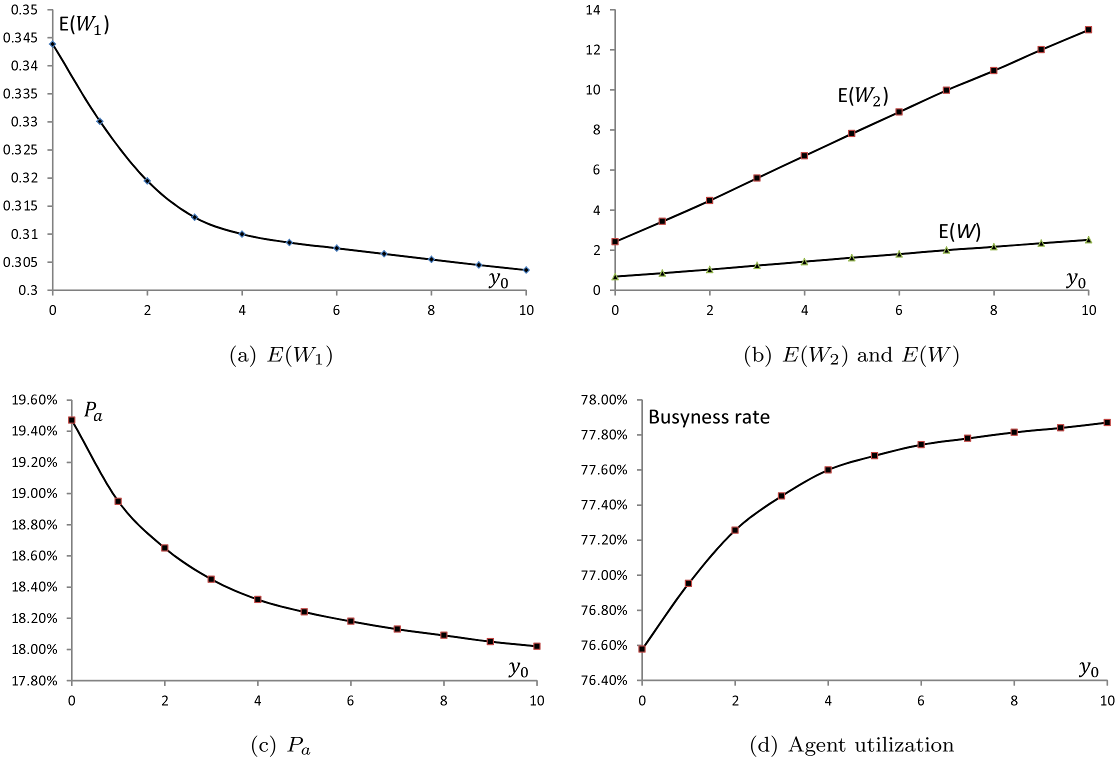
## 5.2 Impact of Agent Reservation

We examine here the impact of the reservation on the system performance. We first consider the two-server case in Section 5.2.1, and second the multi-server case in Section 5.2.2.

### 5.2.1 Two-Server Case

The reason for considering the two-server case is to allow the reservation policy to be only dependent on one parameter; the threshold $y_0$ that defines the limit on the length of queue 2 at and above which no agent should be reserved for inbounds. An illustration of the effect of $y_0$ on the performance measures is given for Model C in Figure 9.

The reservation for inbounds increases in $y_0$. Therefore, $E(W_1)$ and $P_a$ decreases in $y_0$ (Figures 9(a) and 9(c)), and $E(W_2)$ increases in $y_0$ (Figure 9(b)). We observe from Figure 9(b) that reservation deteriorates the overall expected waiting time, $E(W)$. This is related to two reasons. The first one is that agent reservation creates unproductive idling situations with one agent idle while queue 2 is not empty. The latter deteriorates the overall performance of the system. The second reason is related to the reduction of balking and abandonment of inbounds. Since reservation induces more availability of agents for inbounds, it reduces the proportion of lost inbounds ($P_a$). Agents have then to treat more tasks (recall that outbounds do not abandon) as shown in Figure 9(d). This deteriorates the overall expected waiting time. It is the negative effect for reducing the proportion of abandonment.



(a) $E(W_1)$

(b) $E(W_2)$ and $E(W)$

(c) $P_a$

(d) Agent utilization

Figure 9: Effect of $y_0$ ($s = 2$, $\lambda = 1.9$, $\mu_1 = \mu_2 = 1$, $\alpha = q = 30\%$, $r_1 = 100\%$, $k = 0$, $\beta = 0$)

### 5.2.2 Impact of the Call Center Size

We investigate the impact of the call center size $s$ on the performance measures and its relation with the reservation policy. Consider Model C with $k = 0$ (optimal $k$, Proposition 5) and let us define $\rho$ as $\rho = \frac{\lambda}{s\mu_1}$. Proposition 4 provides convexity results justifying that reservation policies bring higher improvement in large call centers than in small ones (recall that non-idling is optimal in the single-server case). For large call centers, these results support therefore the well known notion that only limited server pooling/flexibility/availability is needed (Bassamboo et al., 2010; Legros et al., 2015a).

**Proposition 4** *Consider the non-idling case of Model C. For the optimal threshold on queue 1 ($k = 0$), $P_b$, $\Psi$, $E(W_1)$ and $E(W)$ are decreasing and convex in $s$, when $\rho = a/s$ and $\lambda$ are held constant.*

The proof of the proposition is given in Appendix E. Table 2 illustrates for Model C the behavior of the performance measures as a function of $s$, when $\rho$ is held constant and equal to 0.99. In the second and third columns, we give the upper and lower bounds of $E(W_1)$ using the results of Section 4.3.1, respectively. The upper bound is obtained in the non-idling case and the lower bound is obtained in the highest reservation case. In the fourth column, we compute the relative difference between the upper and lower bounds of $E(W_1)$ so as to assess the possibilities of performance improvement for inbounds. We observe that the higher is $s$, the more it is possible to improve $E(W_1)$. In the fifth column, we give the lower bound of $E(W_2)$ obtained in the non-idling case. We do not give upper bounds of $E(W_2)$ from the extreme reservation case, since these are too high and do not provide interesting situations for the optimization problem. In the sixth column, we give the total expected system cost in the non-idling case. We observe that the expected total cost decreases in $s$. In the seventh, eight and ninth columns, we give the optimal performance measures obtained via the algorithm proposed in Section 4.1, under the optimal reservation policy. In the last column, we compute the relative difference between the optimal system cost and that obtained in the non-idling case. We observe that the larger is the call center, the higher is the agent reservation for inbounds. For small call centers (for $s \leq 5$ in Table 2), non-idling is optimal. As $s$ increases, we observe for the optimal reservation policy that $E(W_1)$ moves from the neighborhood of its upper bound to that of its lower bound, and that $E(W_2)$ remains relatively close to its lower bound. This implies that the relative difference between the system cost in the optimal case and that in the non-idling case increases in the call center size.

Table 2: Effect of $s$ ($\rho = 0.99$, $\mu_1 = \mu_2 = 1$, $r_1 = 1$, $\alpha = \beta = 0$, $q = 30\%$, $k = 0$, $SC = E(W_1) + 0.01E(W_2)$)

| $s$ | $E(W_1)_{\max}$ | $E(W_1)_{\min}$ | rd | $E(W_2)_{\min}$ | $SC_{NI}$ | $E(W_1)_{op}$ | $E(W_2)_{opt}$ | $SC_{opt}$ | rd |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.178 | 3.178 | 0% | 289.36 | 6.07 | 3.178 | 289.36 | 6.07 | 0.00% |
| 2 | 1.584 | 1.367 | -14% | 145.57 | 3.04 | 1.584 | 145.57 | 3.04 | 0.00% |
| 5 | 0.628 | 0.510 | -19% | 68.16 | 1.31 | 0.628 | 68.16 | 1.31 | 0.00% |
| 10 | 0.304 | 0.170 | -44% | 26.09 | 0.56 | 0.250 | 29.04 | 0.54 | -4.27% |
| 20 | 0.150 | 0.063 | -60% | 13.17 | 0.28 | 0.098 | 15.12 | 0.25 | -11.48% |
| 50 | 0.058 | 0.018 | -69% | 6.39 | 0.12 | 0.029 | 6.45 | 0.09 | -23.28% |
| 100 | 0.028 | 0.005 | -82% | 3.93 | 0.07 | 0.009 | 4.04 | 0.05 | -26.11% |
| 500 | 0.004 | 0.001 | -89% | 0.70 | 0.01 | 0.001 | 0.71 | 0.01 | -29.22% |

In summary, the main conclusion of this section is that reservation has more potential of improvement in large call centers, since large call centers allow for less flexibility than small ones.

## 5.3 Impact of the Threshold $k$

We examine the optimization of the threshold $k$. We also investigate the relation between reservation and $k$. Finally, the policy of a fixed threshold $k$ is evaluated in comparison with a state-dependent $k$.

### 5.3.1 Exogenous Parameters and Threshold $k$

Proposition 5 gives, for the non-idling case for Model C, first order monotonicity results in $k$.

**Proposition 5** *In the non-idling case for Model C, $P_b$ is insensitive to $k$, $\Psi$ is decreasing in $k$, $E(W_1)$ and $E(W_2)$ are increasing in $k$, for $k \geq 0$.*

The proof of the proposition is given in Appendix F. A consequence of the monotonicity results of $E(W_1)$ and $E(W_2)$ is that $k = 0$ is optimal for non-idling Model C. Yet, $k = 0$ is not the optimal value for Models A, B and G because of inbounds balking, abandonment and/or the possible non-availability of a called back customer.

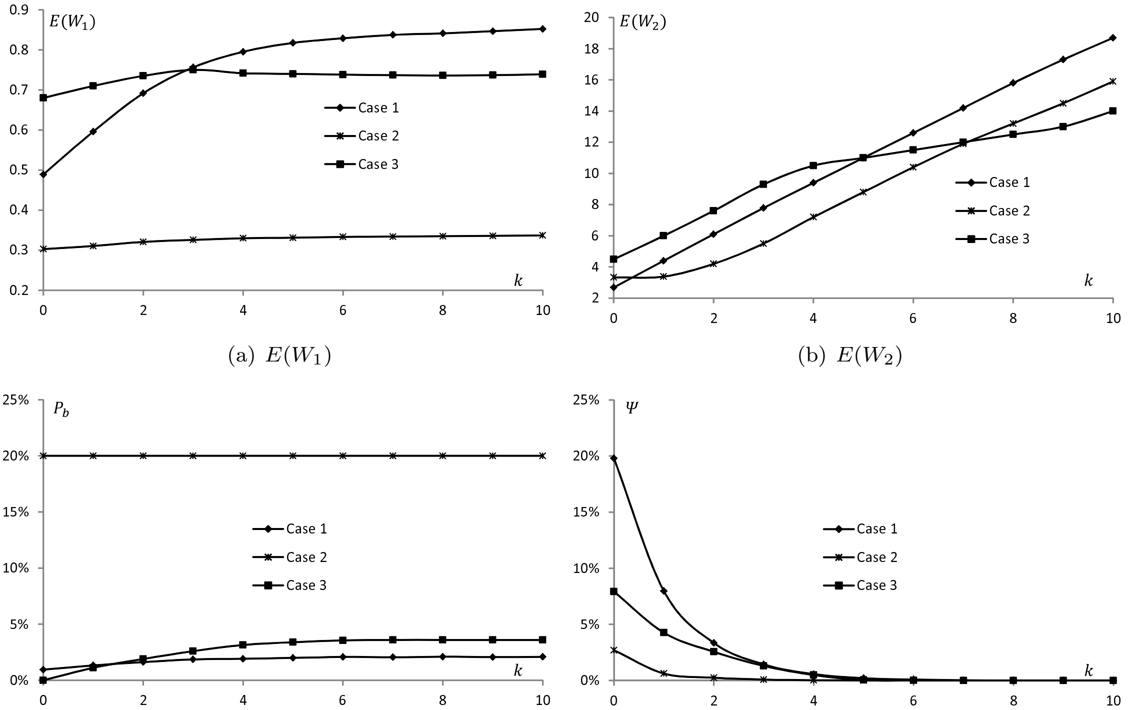
**Balking and call acceptance parameters for Model B.** We consider the impact of $\alpha_x$ and $q_x$ ($x \geq s$) on the monotonicity of the performance measures in $k$ for Model B. In Figure 10, we consider three numerical cases:

- Case 1: $q_x = 0.4$ and $\alpha_x = \min(0.5, 0.05x)$,
- Case 2: $q_x = \min(0.4, 0.05x)$ and $\alpha_x = 0.5$,

- Case 3: $q_x = 0.1$ and $\alpha_x = \min(0.5, 0.05x^2)$,

for $x \geq s$. Cases 1 and 3 illustrate situations with non constant balking parameters and Case 2 illustrates a situation with non constant callback acceptance parameters. The monotonicity results in $k$ in Cases 1 and 2 are identical to those derived for the non-idling case of Model C. However, in Case 3, $E(W_1)$ is non-increasing in $k$.

When $\alpha_x$ is strongly increasing in $x$ (Case 3), the inbounds expected waiting time can be non-increasing in $k$ (Figure 10(a)). The proportion of inbounds increases in $k$. Therefore, inbounds arrive more often at a long queue 1 (large values of $x$), and the balking would then be more important (large values of $\alpha_x$). Although increasing $k$ has the negative effect of increasing balking, it also has the positive effect of reducing the system workload by reducing arrivals that enter the system. This can improve the expected waiting time of inbounds. From the numerical experiments, we however observe that $q_x$ do not impact the first order monotonicity results in $k$. This is related to the fact that $q_x$ has no effect on the system workload, and that the callback offer is only proposed for $x \geq k$.



(a) $E(W_1)$

(b) $E(W_2)$

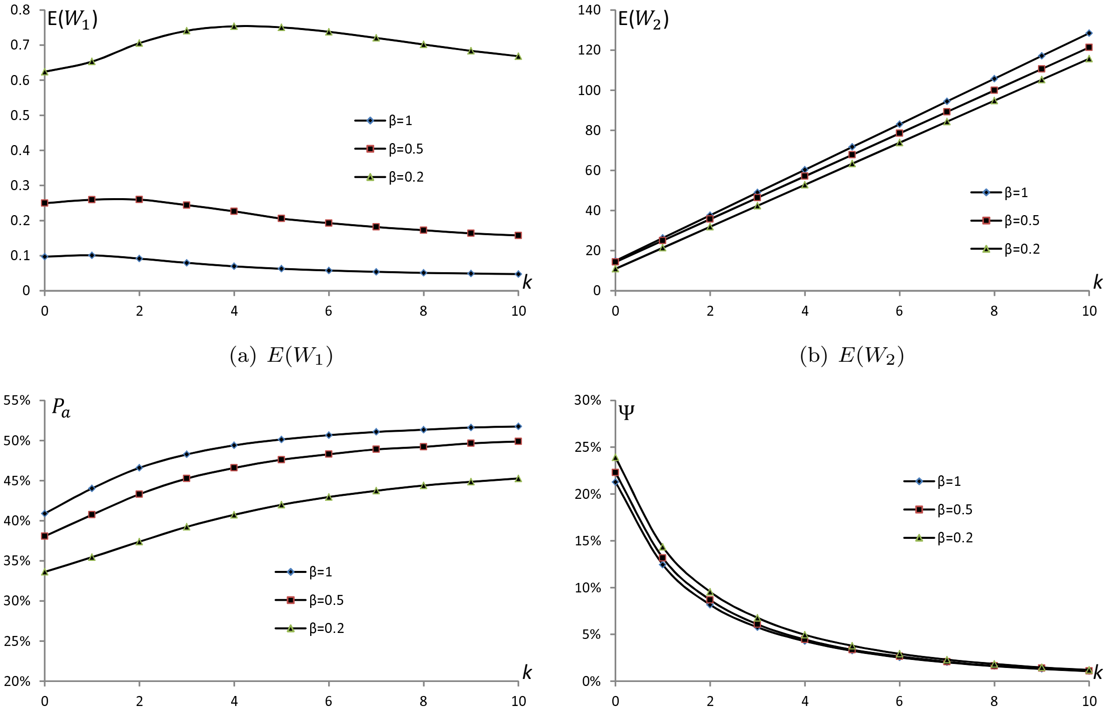
(c) $P_b$

(d) $\Psi$

Figure 10: Impact of balking and call acceptance parameters ($s = 1$, $\lambda = 0.5$, $\mu_1 = 1$, $r_1 = 100\%$, non-idling case)

**Abandonment for Model A.** Figures 11 and 12 illustrate the impact of $k$ on the performance measures, for different values of the abandonment rate $\beta$. We observe that the abandonment in queue 1 only affects the monotonicity properties of $E(W_1)$ and $E(W_2)$. This explains why $k = 0$ is no longer necessarily optimal. Two phenomenons are in competition when $\beta > 0$. From the one hand, increasing $k$ reduces the number of callbacks and increases thus the proportion of inbounds, which would in turn increase $E(W_1)$ and $E(W_2)$.

From the other hand, the increasing of the number of customers in queue 1 increases also the departure rate (after abandonment or service) of inbounds from the system, which makes the system more efficient and may decrease $E(W_1)$ and $E(W_2)$. The first (second) phenomenon is predominant for small (large) values of $\beta$. We observe that the non-increasing of $E(W_2)$ requires higher arrival or abandonment rates than the non-increasing of $E(W_1)$ (Figure 12). The behavior of the other performance measures is more intuitive; the proportion of abandonment increases in $k$, and the proportion of callbacks decreases in $k$.



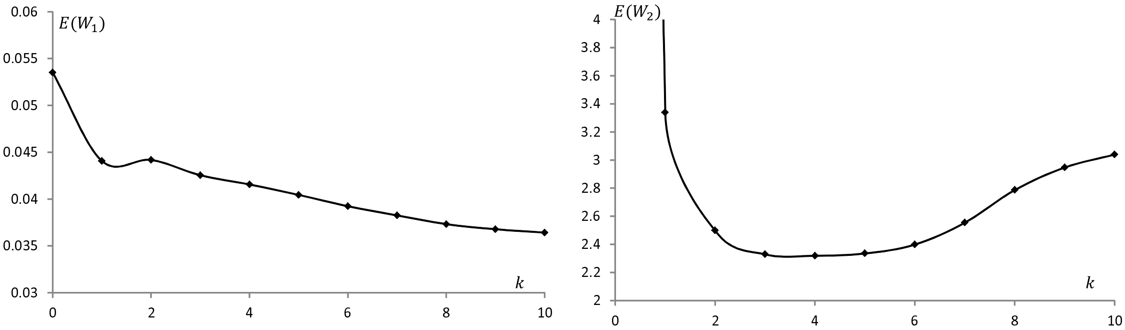
(a) $E(W_1)$

(b) $E(W_2)$

(c) $P_a$

(d) $\Psi$

Figure 11: Impact of abandonment ($s = 1$, $\lambda = 1.2$, $\mu_1 = 1$, $\alpha = q = 30\%$, $r_1 = 100\%$)



(a) $E(W_1)$

(b) $E(W_2)$

Figure 12: Impact of abandonment ($s = 10$, $\lambda = 12$, $\mu_1 = 1$, $\beta = 3$, $\alpha = 10\%$, $q = 30\%$, $r_1 = 100\%$)

**Outbound service process for Model G.** Let us define $T$, a random variable, representing the total time spent by the system capacity to serve an outbound call. For a given outbound call, this corresponds to the summation of the durations spent by agents to handle eventually its non-availability situations plus its service duration.

The case $k = 0$ is also not necessarily optimal when the overall expected time spent to serve an outbound call is larger than the expected time to serve an inbound one. In Proposition 6, we give the expected value and the standard deviation of the time spent by the system capacity to serve an outbound call.

**Proposition 6** *The random time $T$ has a phase type distribution with expected value*

$$E(T) = \frac{r_1}{r_1 + r_2}\frac{1}{\mu_1} + \frac{r_2}{r_1 + r_2}\frac{1}{\mu_2} + \frac{1 - r_1 - r_2}{r_1 + r_2}\frac{1}{\mu_3},$$

*and standard deviation*

$$\sigma(T) = \sqrt{\frac{r_1(4 - 3r_1 - 2r_2)}{\mu_1^2(r_1 + r_2)(2 - r_1 - r_2)} + \frac{r_2(4 - 3r_2 - 2r_1)}{\mu_2^2(r_1 + r_2)(2 - r_1 - r_2)} + \frac{(1 - r_1 - r_2)(4 - r_1 - r_2)}{\mu_3^2(r_1 + r_2)(2 - r_1 - r_2)}}.$$

The proof of this proposition is given in Appendix G. From Proposition 6, we deduce that outbounds require a larger expected time of treatment than that of inbounds if and only if

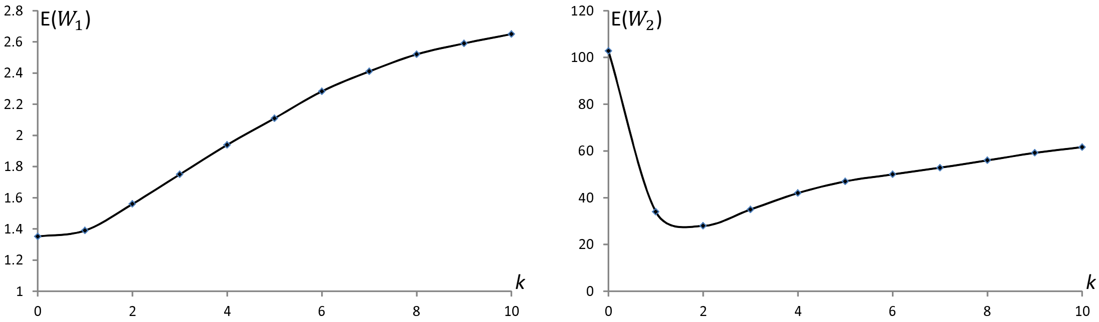
$$\frac{1 - r_1 - r_2}{\mu_3} > r_2\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right). \tag{3}$$

Inequality (3) simply states that if the time lost in handling a non-available situation is larger than the time saved due to fast outbounds (those who have already resolved a part of their problem), then outbounds require a larger expected time of treatment.

Figure 13 illustrates a situation where the overall expected time of an outbound treatment, $E(T)$, is larger than that of the service time of an inbound, $1/\mu_1$. We observe that the monotonicity properties in $k$ of the performance measures $E(W_1)$, $P_a$ and $\Psi$ are not affected by the parameters of service of outbounds, because of the higher priority given to inbounds. The reason is that, during their sojourn in the queue, the latter will only assist at service durations that are exponentially distributed with rate $\mu_i$ ($i = 1, 2, 3$). We observe that $E(W_2)$ is either strictly increasing in $k$ or decreasing then increasing. The second situation occurs when outbounds are treated within a much larger time than that of inbounds. Two phenomenons are in competition; the first one already mentioned earlier is that increasing $k$ reduces the number of outbounds which would suffer from the high proportion of prioritized inbounds. The second one is that if $k$ is too small, the proportion of outbounds can be too important for the system capacity. It might then take too long time to serve them.

### 5.3.2 Reservation and Threshold $k$

We investigate here the relation between the agent reservation policy and the choice for the threshold $k$. We proved in Proposition 2 that for Model A, the higher is $q$, the less agents should be reserved for inbounds. The reason is the low proportion of inbounds. The impact of $k$ is similar to that of $q$. Increasing $k$ is equivalent to decreasing $q$, therefore the higher $k$ is the more agents should be reserved for inbounds. This observation agrees with the classical idea in control problems stating that the longest queue should be preferred: through the choice of the reservation level in our model.



(a) $E(W_1)$

(b) $E(W_2)$

Figure 13: Impact of the service process ($s = 1$, $\lambda = 0.75$, $\mu_1 = 1$, $\mu_2 = 1.5$, $\mu_3 = 10$, $\alpha = \beta = 0$, $q = 30\%$, $r_1 = 0\%$, $r_2 = 7\%$)

However, Table 3 reveals that this observation is no longer true when Model G is considered. In this table, we provide the performance measures for different values of $k$. Similarly to Table 2, we provide the upper and lower bounds for $E(W_1)$ and $P_a$ to examine the possibilities of improvement. We also compute the lower bound for $E(W_2)$. In the presented numerical illustration, the two extreme situations are again the non-idling case and the extreme reservation case. In the last five columns, we give the optimal values of the performance measures. We also compute the relative difference found in the comparison between the non-idling case and the optimal case.

On the contrary to what one would expect, we observe here that agent reservation decreases in $k$. For example in Table 3, when $k \geq 6$, non-idling is optimal. The reason is related to two phenomenons. The first one is the possible non-availability of outbounds (20% are not available). The second one is the smaller impact of outbounds in service on inbounds performance when $r_1 < 1$ than when $r_1 = 1$ (50% of outbounds occupy agents a shorter time than inbounds). The low priority of outbounds together with their non-full availability make queue 2 difficult to reduce, especially when inbounds are numerous in the system (i.e., when $k$ is high). Therefore, the increasing of $E(W_2)$ in $k$ is strong (see column 8) and reservation for inbounds should not be provided when $k$ is high. Because outbounds occupy agents a shorter time than inbounds when $r_1 < 1$, outbounds have less impact on $E(W_1)$ in Model G than in Model A. Thus, the effect of $k$ and the agent reservation on $E(W_1)$ is weaker for Model G than for Model A (see columns 2 and 3). Increasing reservation when $k$ is high has a strong impact on $E(W_2)$ but a small one on $E(W_1)$, which advocates for a non-idling policy. The deterioration of $E(W_2)$ with reservation is weaker when $k$ is small, so, reservation should be provided in this case to reduce $E(W_1)$.

### 5.3.3 Value of a Fixed Threshold $k$

We have defined the threshold parameter $k$ on the number of calls in queue 1 to control the decision of proposing or not the callback offer. We have shown that $k = 0$ is optimal for the non-idling case of Model C. In other words, the callback offer should be proposed to all delayed customers. It is also the case for Models A, B and G in most cases. Yet, with significant abandonment or large treatment time for outbounds, $k = 0$ may not be any longer optimal. In the modeling, the value of a fixed threshold $k$ comes from its simplicity

Table 3: Impact of $k$ ($\lambda = 49.5$, $s = 50$, $\mu_1 = 1$, $\mu_2 = 1.5$, $\mu_3 = 10$, $r_1 = 50\%$, $r_2 = 30\%$, $\alpha = 10\%$, $q = 30\%$, $\beta = 0.5$, $SC = E(W_1) + 0.01E(W_2) + P_a$)

| $k$ | $E(W_1)_{\max}$ | $E(W_1)_{\min}$ | rd | $P_{a_{\max}}$ | $P_{a_{\min}}$ | rd | $E(W_2)$ | $SC_{NI}$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.013 | 0.010 | -21.16% | 4.85% | 4.04% | -20.13% | 0.343 | 0.065 |
| 1 | 0.015 | 0.012 | -19.68% | 5.02% | 4.20% | -19.62% | 0.497 | 0.070 |
| 2 | 0.018 | 0.015 | -18.81% | 5.21% | 4.44% | -17.50% | 0.656 | 0.076 |
| 3 | 0.020 | 0.017 | -18.34% | 5.32% | 4.56% | -16.87% | 0.799 | 0.081 |
| 4 | 0.023 | 0.019 | -17.76% | 5.43% | 4.68% | -16.17% | 0.918 | 0.086 |
| 5 | 0.025 | 0.021 | -17.34% | 5.52% | 4.77% | -15.80% | 1.039 | 0.090 |
| 6 | 0.027 | 0.023 | -17.18% | 5.56% | 4.82% | -15.22% | 1.158 | 0.094 |
| 7 | 0.029 | 0.025 | -15.65% | 5.69% | 4.96% | -14.78% | 1.247 | 0.098 |
| 8 | 0.031 | 0.027 | -12.55% | 5.74% | 5.08% | -13.06% | 1.349 | 0.102 |
| 9 | 0.032 | 0.029 | -9.63% | 5.82% | 5.28% | -10.23% | 1.444 | 0.105 |

| $k$ | $E(W_1)_{opt}$ | $E(W_2)_{opt}$ | $P_{a_{opt}}$ | $SC_{opt}$ | rd |
|---|---|---|---|---|---|
| 0 | 0.011 | 0.456 | 4.12% | 0.057 | -12.07% |
| 1 | 0.014 | 0.594 | 4.26% | 0.063 | -10.69% |
| 2 | 0.017 | 0.687 | 4.48% | 0.068 | -10.44% |
| 3 | 0.019 | 0.861 | 4.63% | 0.074 | -9.72% |
| 4 | 0.021 | 0.963 | 4.80% | 0.079 | -8.84% |
| 5 | 0.024 | 1.087 | 4.99% | 0.084 | -6.83% |
| 6 | 0.027 | 1.158 | 5.56% | 0.094 | 0.00% |
| 7 | 0.029 | 1.247 | 5.69% | 0.098 | 0.00% |
| 8 | 0.031 | 1.349 | 5.74% | 0.102 | 0.00% |
| 9 | 0.032 | 1.444 | 5.82% | 0.105 | 0.00% |

and from the analysis tractability for the performance evaluation. However, a fixed threshold $k$ may not be optimal. It is then also interesting to evaluate the performance of our fixed-$k$ policy in comparison with a state-dependent-$k$ policy for the proposition of the callback offer upon arrival.

For Model A with constant balking and callback acceptance parameters, the value functions defined in Section 2 can be rewritten, for $n \geq 0$, including the decision to propose or not the callback offer through the operator $W_n$, as

$$U_{n+1}(x,y) = \gamma_1(x-s)^+ + \gamma_2 y + \lambda W_n(x,y) + \beta(x-s)^+(V_n(x-1,y)+\gamma_3) + \min(s,x)\mu V_n(x-1,y)$$
$$+ \left(1 - \lambda - \beta(x-s)^+ - \min(s,x)\mu_1\right) V_n(x,y), \text{ for } x, y \geq 0,$$

with

$$V_{n+1}(x,y) = \min(U_{n+1}(x+1,y-1), U_{n+1}(x,y)),$$

for $y > 0$ and $0 \leq x < s$ and $V_{n+1}(x,y) = U_{n+1}(x,y)$ in the remaining cases, and

$$W_{n+1}(x,y) = \min(\alpha(U_{n+1}(x,y)+\gamma_3) + (1-\alpha)U_{n+1}(x+1,y), \alpha(U_{n+1}(x,y)+\gamma_3) + qU_{n+1}(x,y+1) + (1-q-\alpha)U_{n+1}(x+1,y)),$$

for $x \geq s$ and $W_{n+1}(x,y) = U_{n+1}(x,y)$ in the remaining cases. We choose $W_0(x,y) = V_0(x,y) = U_0(x,y) = 0$, for $x, y \geq 0$.

In Figure 14, we present the optimal decision found through value iterations. We only present the states where an action on the callback offer has to be taken ($x \geq s$). We observe that the optimal decision for the callback offer if of switch type. The optimal decision for the points on the curve is not to propose the callback offer. To the contrary to the reservation policy found in Section 3, the switching curve is not monotonous in $x$ or in $y$.

We observe that if the optimal decision in a given state $(x, y)$ is not to propose the callback offer, then the same decision should be taken in state $(x, y+1)$. The reason is related to the congestion of queue 2. The decision not to propose the offer is taken in order to use call abandonment in queue 1 which decreases the system workload. Therefore, if the system is too congested with $y$ outbounds in queue 2, it would also be

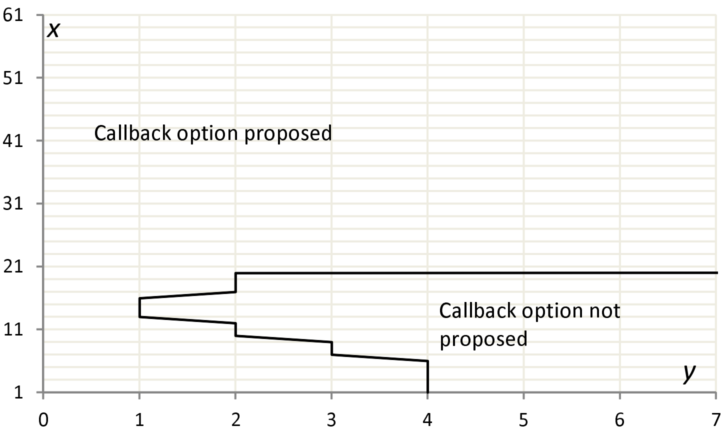


Figure 14: Optimal switching curve for the callback offer ($\lambda = 1.2$, $\mu_1 = 1$, $q = \alpha = 30\%$, $\beta = 1$, $SC = E(W_1) + 0.0005E(W_2) + 0.1P_a$, $r_1 = 1$, $s = 1$)

with $y+1$ outbounds in queue 2. The decision as a function of $x$ is more complex. For small values of $x$, the decision is more likely to give the offer so as to reduce the number of customers in queue 1. This decision can be taken because the system is not congested. For higher values of $x$, the offer can be interrupted to reduce the workload in the system by letting customers abandon from queue 1. For even higher values of $x$, the proportion of abandonment and the waiting time in queue 1 can be so significant that the decision is again to propose the callback offer even if it would increase the system workload.

One can compare between the two modelings, with a fixed or a state-dependent $k$ using simulations. The optimal fixed-$k$ is assumed to be a real number in order to achieve a lower system cost than if $k$ would be an integer (in practice, this means that randomization between two adjacent thresholds is allowed). For various settings, Table 4 reveals that the difference between the optimal system cost and the cost found with a fixed-$k$ is not important. However, it is notable that the optimal state-dependent policy improves $E(W_2)$ and almost do not affect the other metrics.

Table 4: Comparison between the two threshold modelings ($\mu_1 = 1$, $r_1 = 1$)

| Cases | Parameters | | | | | |
|---|---|---|---|---|---|---|
| | $\lambda$ | $s$ | $\beta$ | $\alpha$ | $q$ | $SC$ |
| 1 | 1.2 | 1 | 1 | 30% | 30% | $E(W_1) + 0.0005E(W_2) + 0.1P_a$ |
| 2 | 0.99 | 1 | 2 | 0% | 80% | $E(W_1) + 0.00004E(W_2) + 0.01P_a$ |
| 3 | 0.8 | 1 | 1.5 | 10% | 40% | $E(W_1) + 0.0001E(W_2) + 0.01P_a$ |
| 4 | 12 | 10 | 3 | 10% | 60% | $E(W_1) + 0.0001E(W_2) + 0.2P_a$ |
| 5 | 120 | 100 | 3 | 10% | 60% | $E(W_1) + 0.0000001E(W_2)$ |

| Cases | Optimal fixed-$k$ policy | | | | | Optimal state-dependent-$k$ policy | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $k$ | $E(W_1)$ | $E(W_2)$ | $P_a$ | $SC$ | $E(W_1)$ | $E(W_2)$ | $P_a$ | $SC$ |
| 1 | 3.1 | 0.095 | 39.200 | 46.8% | 0.161 | 0.095 | 36.100 | 46.8% | 0.160 |
| 2 | 3.9 | 0.020 | 897.980 | 42.2% | 0.060 | 0.021 | 464.260 | 41.3% | 0.044 |
| 3 | 6.8 | 0.030 | 61.878 | 40.7% | 0.041 | 0.031 | 35.689 | 40.8% | 0.038 |
| 4 | 1.2 | 0.044 | 3.338 | 20.1% | 0.085 | 0.044 | 3.338 | 20.1% | 0.085 |
| 5 | 0.9 | 0.014 | 580.393 | 12.0% | 0.014 | 0.014 | 542.040 | 12.0% | 0.014 |

For Model G, the optimal decisions for the callback offer can be obtained using the same approach. However, further assumptions should be made on the balking parameters when the callback option would be proposed or not. For instance, it seems appropriate to assume that the callback offer would reduce the balking behavior. In this case, the conclusion derived above are still valid. The callback offer reduces then at the same time balking, abandonment and the waiting time in queue 1, but it increases the system workload. For Model G, either the treatment time of outbounds is shorter than that of inbounds and $k = 0$ is thus optimal, or it is not and the conclusions derived above are also still valid.

To conclude Section 5.3, $k = 0$ is optimal when the balking parameters are constant ($\alpha_x = \alpha$, $x \geq s$), no abandonment is considered ($\beta = 0$), or the treatment time of an outbound call is lower or equal than that of an inbound one ($E(T) \leq 1/\mu_1$). Increasing $k$ increases the size of queue 1. When $\alpha_x$ is strongly increasing in $x$, this also increases the balking proportion which reduces the effective arrival rate. When $\beta > 0$, call abandonment helps to reduce the length of queue 1. If much more importance is given to the waiting time in queue 1 than that to abandonment ($\gamma_1 >> \gamma_3$), then $k > 0$ is useful to discharge the system. If the treatment time of an outbound call is large, it is also useful to have $k > 0$ in order to avoid too high proportion of outbounds. The relation between the optimal reservation policy and the optimal $k$ depends on the service process of outbounds. If this one is identical to that of inbounds (Model A), then more agents should be reserved for inbounds as $k$ increases.

**Summary of Section 5 results.** Table 5 summaries the impact of the parameters on the objective function components. We use the sign "+" for a positive effect and the sign "-" for a negative one.

Table 5: Impact of the parameters

| | $E(W_1)$ | $E(W_2)$ | $P_a$ |
|---|---|---|---|
| Increasing $s$ | + | + | + |
| Increasing the agent reservation | + | - | + |
| Increasing $k$ | - | - | - |
| Increasing $k$ with a high balking/abandonment parameters | + | +,- | - |
| Increasing $k$ with a high difficulty to serve outbounds | - | +,- | - |

In most observed situations, $k = 0$ is optimal and the reservation policy can be obtained via the MDP approach from Section 3. In the remaining cases, a finite number of steps should be done to find the optimal value of $k$ with its corresponding reservation policy (by starting from the case $k = 0$ and by incrementing $k$ by one at each step). The number of tests is finite because the deterioration of $E(W_2)$ in $k$ after a given value of $k$ is much faster than the eventual improvement of $E(W_1)$ in $k$. Beyond this value of $k$, any reservation policy would anyway further deteriorate $E(W_2)$. Moreover, $P_a$ deteriorates with $k$. Hence, after a given value of $k$, the total expected system cost only increases in $k$ and the search for the optimal value of $k$ should be stopped at that point.

# 6 Conclusions and Future Research

We considered a call center that offers two channels: real-time telephone service and postponed (callback) service. Customers choose which channel to use based on a probabilistic choice model. We demonstrated the operational advantages of agent reservation in this context.

The key operational findings of this paper are that (1) the value of the callback option is more significant under heavily loaded situations, (2) the benefits of agent reservation are more apparent in large call centers than in small ones, (3) reservation increases the agent utilization due to the abandonment reduction, (4) reservation is not likely to be used under light or heavily loaded situations, (5) the callback offer should be proposed to all delayed customers except when the abandonment is significant or when the overall treatment time of an outbound is much larger than that of an inbound. These operational findings came together with theoretical contributions. The major ones are (1) the proof that non-idling is optimal in the single-server case, (2) the proof of the optimality of a threshold policy in the two-server case, (3) the algorithm proposed for the performance evaluation when transition rates are assumed to be constant.

Several interesting areas of future research arise. It would be useful to empirically validate the customer reaction model to the callback offer through real data analysis. It is also interesting to extend the proof of the optimal policy for the two-server case to that for the multi-server case. Another research avenue is to consider other optimization problem formulations, for example in terms of quantiles on the waiting time distributions of inbound and outbound calls. Finally, it would be interesting to consider non-stationary arrival parameters and investigate its impact on job scheduling.

# Acknowledgements

This work was supported by Agence Nationale de la Recherche under the project ANR-JCJC-SIMI3-2012-OPERA. We also want to express our gratitude to two anonymous reviewers and the associate editor for their useful comments, that significantly improved this 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*, 16:665–688.

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

Armony, M. and Maglaras, C. (2004b). On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Operations Research*, 52(2):271–292.

Armony, M. and Ward, A. (2010). Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, 58(3):624–637.

Bassamboo, A., Randhawa, R., and van Mieghem, J. (2010). Optimal flexibility configurations in newsvendor networks: Going beyond chaining and pairing. *Management Science*, 56:1285–1303.

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

Bhulai, S., Brooms, A., and Spieksma, F. (2014). On structural properties of the value function for an unbounded jump Markov process with an application to a processor sharing retrial queue. *Queueing Systems*, 76:425–446.

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

de Véricourt, F. and Zhou, Y.-P. (2005). Managing response time in a call routing problem with service failure. *Operations Research*, 53:968–981.

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*, 34:1616–1645.

Dudin, S., Kim, C., Dudina, O., and Baek, J. (2013). Queueing system with heterogeneous customers as a model of a call center with a call-back for lost customers. *Mathematical Problems in Engineering.* ID 983723, Volume 2013, 13 pages.

Dumas, G. P., Perkins, M. M., and White, C. M. (1996). Call sharing for inbound and outbound call center agents. US Patent 5,519,773.

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

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

Gurvich, I., Armony, M., and Maglaras, C. (2009). Cross-selling in a call center with a heterogeneous customer population. *Operations research*, 57(2):299–313.

Hajek, B. (1984). Optimal control of two interacting service stations. *Automatic Control, IEEE Transactions on*, 29(6):491–499.

Harel, A. (2011). Convexity results for the erlang delay and loss formulae when the server utilization is held constant. *Operations research*, 59(6):1420–1426.

ICMI (2013). Extreme engagement in the multichannel contact center: Leveraging the emerging channels research Report and best practices guide. ICMI Research Report.

Kim, C., Dudina, O., Dudin, A., and Dudin, S. (2012). Queueing system MAP/M/N as a model of call center with call-back option. *Chapter in Analytical and Stochastic Modeling Techniques and Applications, Series Lecture Notes in Computer Science, Springer Berlin Heidelberg, Editors: Al-Begain, K. and Fiems, D. and Vincent, J.M.*, 7314:1–15.

Koole, G. (2007). *Monotonicity in Markov reward and decision chains: Theory and applications*, volume 1. Now Publishers Inc.

Koole, G. (2013). *Call Center Optimization.* MG Books.

Legros, B., Jouini, O., and Dallery, Y. (2015a). A flexible architecture for call centers with skill-based routing. *International Journal of Production Economics*, 159:192–207.

Legros, B., Jouini, O., and Koole, G. (2013). Front-line service employee multi-tasking in the presence of customer self-service tasks. Working paper. Ecole Centrale Paris.

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

Lin, W. and Kumar, P. (1984). Optimal control of a queueing system with two heterogeneous servers. *Automatic Control, IEEE Transactions on*, 29(8):696–703.

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.

Puterman, M. (1994). *Markov Decision Processes.* John Wiley and Sons.

Queffélec, H. and Zuily, C. (2013). *Analyse pour l'Agrégation.* Collection: Sciences Sup, Dunod, Paris.

# Appendix

## A Proof of Theorem 1

We first rewrite the value functions in the two-server case for Model A ($\mu_1 = \mu_2 = \mu$, $r_1 = 1$). So as to simplify the presentation of the proof, we redefine the states as follows. The parameter $z$ denotes the state of the agents team ($z = 0$ when both agents are idle; $z = 1$ when only one agent is busy with an inbound or an outbound call; and $z = 2$ when both agents are busy), $x$ is redefined here as the number of inbounds in queue 1 and $y$ is the number of outbounds in queue 2. We have for $n \geq 0$,

$$\begin{aligned}
U_{n+1}(0,0,y) &= \gamma_2 y + \lambda V_n(1,0,y) + (1-\lambda)V_n(0,0,y), \text{ for } y \geq 0,\\
U_{n+1}(1,0,y) &= \gamma_2 y + \lambda V_n(2,0,y) + \mu V_n(0,0,y) + (1-\lambda-\mu)V_n(1,0,y), \text{ for } y \geq 0,\\
U_{n+1}(2,x,y) &= \gamma_1 x + \gamma_2 y + \lambda \big(\mathbf{1}_{(0\leq x<k)}\left((1-\alpha_x)V_n(2,x+1,y) + \alpha_x(V_n(2,x,y)+\gamma_3)\right)\\
&\quad +\mathbf{1}_{(k\leq x<N)}\left(q_x V_n(2,x,y+1) + \alpha_x(V_n(2,x,y)+\gamma_3) + (1-q_x-\alpha_x)V_n(2,x+1,y)\right)\\
&\quad +\mathbf{1}_{(x=N)}(q_{N-1}V_n(2,x,y+1) + (1-q_{N-1})(V_n(2,x,y)+\gamma_3))\big)\\
&\quad + \beta x(V_n(2,x-1,y)+\gamma_3) + 2\mu\left(\mathbf{1}_{(x=0)}V_n(1,0,y) + \mathbf{1}_{(x>0)}V_n(2,x-1,y)\right)\\
&\quad + (1-\lambda-\beta x-2\mu)\,V_n(2,x,y), \text{ for } x,y \geq 0,
\end{aligned}$$

with $V_{n+1}(0,0,y) = U_{n+1}(1,0,y-1)$ for $y > 0$ (recall that we assume that it is optimal to serve an outbound call when all agents are idling); $V_{n+1}(1,0,y) = \min(U_{n+1}(2,0,y-1), U_{n+1}(1,0,y))$ for $y > 0$; and $V_{n+1}(z,x,y) = U_{n+1}(z,x,y)$ in the remaining cases. We choose $V_0(z,x,y) = U_0(z,x,y) = 0$, for $z = 0,1,2$ and $x,y \geq 0$.

We define a class of functions $\mathcal{F}$ from $\{0,1,2\} \times \mathbb{N}^2$ to $\mathbb{R}$ as follows: $f \in \mathcal{F}$ if

$$f(2,x+1,y) \geq f(2,x,y), \tag{4}$$

$$f(1,0,y) \geq f(0,0,y), \tag{5}$$

$$f(2,0,y) \geq f(1,0,y), \tag{6}$$

$$f(2,x,y+1) \geq f(2,x,y), \tag{7}$$

$$f(0,0,y+1) \geq f(0,0,y), \tag{8}$$

$$f(1,0,y+1) \geq f(1,0,y), \tag{9}$$

$$f(2,x,y) + f(2,x+1,y+1) \geq f(2,x+1,y) + f(2,x,y+1), \tag{10}$$

$$f(0,0,y) + f(1,0,y+1) \geq f(1,0,y) + f(0,0,y+1), \tag{11}$$

$$f(1,0,y) + f(2,0,y+1) \geq f(2,0,y) + f(1,0,y+1), \tag{12}$$

$$f(2,x,y+2) + f(2,x+1,y) \geq f(2,x,y+1) + f(2,x+1,y+1), \tag{13}$$

$$f(0,0,y+2) + f(1,0,y) \geq f(0,0,y+1) + f(1,0,y+1), \tag{14}$$

$$f(1,0,y+2) + f(2,0,y) \geq f(1,0,y+1) + f(2,0,y+1), \tag{15}$$

for $x, y \geq 0$. Relations (4) and (7) define a class of increasing functions in $x$ and in $y$. Relation (10) defines supermodularity for $z = 2$. By summing up Relations (10) and (13) we obtain $f(2, x, y) + f(2, x, y + 2) \geq 2f(2, x, y + 1)$, by summing up Relations (11) and (14) we obtain $f(0, 0, y) + f(0, 0, y + 2) \geq 2f(0, 0, y + 1)$, and by summing up Relations (12) and (15) we obtain $f(1, 0, y) + f(1, 0, y + 2) \geq 2f(1, 0, y + 1)$. Thus if $f \in \mathcal{F}$, then $f$ is convex in $y$. Relation (13) means that the function $f(2, x, y+1) - f(2, x+1, y)$ is increasing in $y$.

**Remark 1** *For the multi-server case of Model G, we need to add another relation to the class of functions defined below. The additional relation is* $f(x+2, y, s_2, s_3) + f(x, y+1, s_2, s_3) \geq f(x+1, y, s_2, s_3) + f(x+1, y+1, s_2, s_3)$. *It is required to prove that the relation* $f(x, y+2, s_2, s_3) + f(x+1, y, s_2, s_3) \geq f(x, y+1, s_2, s_3) + f(x+1, y+1, s_2, s_3)$ *propagates through the minimizing operator. The proof through value iteration is hard to do for the arrival term if* $x = s + k - 2$, *and for the service term if* $0 \leq x + s_2 + s_3 \leq s - 2$. *It is however doable for the remaining cases.*

To simplify the presentation, we denote by "*serve*" the decision action to serve an outbound call, and by "*keep*" the decision action to keep an outbound call in queue 2. The proof of the optimality of the threshold policy reduces to show that Relation (13) is true for $U_n$, $n \geq 0$. We next prove by induction on $n$ that both $V_n$ and $U_n$ are in $\mathcal{F}$. We divide the proof into the following 5 steps:

- **Step 1.** We prove that $V_0, U_0 \in \mathcal{F}$.
- **Step 2.** We prove that if $U_n \in \mathcal{F}$, then $V_n \in \mathcal{F}$, for $n \geq 0$.
- **Step 3.** We prove that the cost term $G(z, x, y) = \gamma_1 x + \gamma_2 y$ is in $\mathcal{F}$.
- **Step 4.** We prove for a given $n \geq 0$ that if $V_n \in \mathcal{F}$, then the following arrival term is also in $\mathcal{F}$:

$$\begin{aligned}A_n(2, x, y) = {} & \mathbf{1}_{(0 \leq x < k)} \left((1 - \alpha_x)V_n(2, x + 1, y) + \alpha_x(V_n(2, x, y) + \gamma_3)\right) \\ & + \mathbf{1}_{(k \leq x < N)} \left(q_x V_n(2, x, y + 1) + \alpha_x(V_n(2, x, y) + \gamma_3) + (1 - q_x - \alpha_x)V_n(2, x + 1, y)\right) \\ & + \mathbf{1}_{(x=N)}(q_{N-1}V_n(2, x, y + 1) + (1 - q_{N-1})(V_n(2, x, y) + \gamma_3)),\end{aligned}$$

  for $x, y \geq 0$, $A_n(1, 0, y) = V_n(2, 0, y)$ and $A_n(0, 0, y) = V_n(1, 0, y)$ for $y \geq 0$.

- **Step 5.** We prove for a given $n \geq 0$ that if $V_n \in \mathcal{F}$, then the following departure term is also in $\mathcal{F}$:

$$\begin{aligned}D_n(2, x, y) = {} & \beta x(V_n(2, x - 1, y) + \gamma_3) + 2\mu \left(\mathbf{1}_{(x=0)}V_n(1, 0, y) + \mathbf{1}_{(x>0)}V_n(2, x - 1, y)\right) \\ & + (1 - \lambda - \beta x - 2\mu)\, V_n(2, x, y),\end{aligned}$$

  for $x, y \geq 0$, $D_n(1, 0, y) = \mu V_n(0, 0, y) + (1 - \lambda - \mu)V_n(1, 0, y)$ and $D_n(0, 0, y) = (1 - \lambda)V_n(0, 0, y)$ for $y \geq 0$.

The proofs for the previous five steps are given below.

**Step 1.** For $x, y \geq 0$ and $z = 0, 1, 2$, $V_0(z, x, y) = U_0(z, x, y) = 0$. Then $V_0, U_0 \in \mathcal{F}$.

**Step 2.** Assume that for a given $n \geq 0$, $U_n \in \mathcal{F}$. We only consider the non-trivial cases where $z = 0$ or $z = 1$ and $y > 0$. In the other cases $U_n = V_n$. Therefore we only need to show Relations (5), (6), (8), (9), (11), (12), (14) and (15).

- For Relations (5) and (8), we have

$$V_n(0, 0, y) = U_n(1, 0, y - 1), \text{ for } y > 0. \tag{16}$$

If "keep" is optimal in $(1, 0, y)$, then $V_n(1, 0, y) = U_n(1, 0, y)$. Combining Equation (16) with Relation (9) for $U_n$ leads to $V_n(0, 0, y) \leq V(1, 0, y)$ and proves Relations (5) for $V_n$. If "serve" is optimal in $(1, 0, y)$, then $V_n(1, 0, y) = U_n(2, 0, y-1)$. Combining Equation (16) with Relation (6) for $U_n$ leads to $V_n(0, 0, y) \leq V(1, 0, y)$ and proves Relations (5) for $V_n$.

We have $V_n(0, 0, y+1) = U_n(1, 0, y)$. Combining Inequality (16) with Relation (9) for $U_n$ leads to $V_n(0, 0, y) \leq V(0, 0, y + 1)$. Therefore in all cases, Relations (5) and (8) hold for $V_n$.

- For Relations (6) and (9), we have

$$V_n(1, 0, y) \leq U_n(2, 0, y - 1), \text{ for } y > 0, \tag{17}$$

$$V_n(1, 0, y) \leq U_n(1, 0, y), \text{ for } y \geq 0. \tag{18}$$

Observe that $V_n(2, 0, y) = U_n(2, 0, y)$. Combining Inequality (17) with Relation (7) for $U_n$ proves Relation (6).

If "keep" is optimal in $(1, 0, y+1)$, then $V_n(1, 0, y+1) = U_n(1, 0, y+1)$. Combining equality (18) with Relation (9) for $U_n$ proves Relation (9) for $V_n$. If "serve" is optimal in $(1, 0, y + 1)$, then $V_n(1, 0, y + 1) = U_n(2, 0, y)$. Combining equality (17) with Relation (7) for $U_n$ proves (9) for $V_n$. Therefore in all cases, Relations (6) and (9) hold for $V_n$.

- For Relation (11), we have

$$V_n(1, 0, y) + V_n(0, 0, y + 1) \leq 2U_n(1, 0, y) \text{ for } y \geq 0, \tag{19}$$

$$V_n(1, 0, y) + V_n(0, 0, y + 1) \leq U_n(2, 0, y - 1) + U_n(1, 0, y) \text{ for } y \geq 0. \tag{20}$$

If "keep" is the optimal action in state $(1, 0, y+1)$, for $y > 0$, then $V_n(0, 0, y) + V_n(1, 0, y+1) = U_n(1, 0, y-1) + U_n(1, 0, y+1)$. Thus, combining the convexity in $y$ of $U_n$ and Inequality (19) proves Relation (11) for $V_n$, for $y \geq 0$. If "serve" is the optimal action in state $(1, 0, y + 1)$, for $y > 0$, then $V_n(0, 0, y) + V_n(1, 0, y + 1) = U_n(1, 0, y-1) + U_n(2, 0, y)$. Combining Relation (12) for $U_n$ and Inequality (20) proves Relation (11) for $V_n$, for $y > 0$. In all cases, Relation (11) then holds for $V_n$.

- For Relation (12), we have

$$V_n(2, 0, y) + V_n(1, 0, y + 1) \leq U_n(2, 0, y) + U_n(1, 0, y + 1) \text{ for } y \geq 0, \tag{21}$$

$$V_n(2, 0, y) + V_n(1, 0, y + 1) \leq 2U_n(2, 0, y) \text{ for } y \geq 0. \tag{22}$$

If "keep" is the optimal action in state $(1,0,y)$, for $y > 0$, then $V_n(1,0,y) + V_n(2,0,y+1) = U_n(1,0,y) + U_n(2,0,y+1)$. Thus, Relation (12) for $U_n$ and Inequality (21) prove Relation (12) for $V_n$, for $y \geq 0$. If "serve" is the optimal action in state $(1,0,y)$, for $y > 0$, then $V_n(1,0,y) + V_n(2,0,y+1) = U_n(2,0,y-1) + U_n(2,0,y+1)$. Combining the convexity in $y$ of $U_n$ and Inequality (22) proves Relation (12) for $V_n$, for $y > 0$. In all cases, Relation (12) then holds for $V_n$.

- For Relation (14), we have

$$V_n(0,0,y+1) + V_n(1,0,y+1) \leq U_n(1,0,y) + U_n(1,0,y+1) \text{ for } y \geq 0, \tag{23}$$

$$V_n(0,0,y+1) + V_n(1,0,y+1) \leq U_n(1,0,y) + U_n(2,0,y) \text{ for } y \geq 0. \tag{24}$$

If "keep" is the optimal action in states $(1,0,y)$, for $y > 0$, then $V_n(0,0,y+2) + V_n(1,0,y) = U_n(1,0,y+1) + U_n(1,0,y)$. Inequality (23) proves Relation (14) for $V_n$, for $y \geq 0$. If "serve" is the optimal action in state $(1,0,y)$, for $y \geq 0$, then $V_n(0,0,y+2) + V_n(1,0,y) = U_n(1,0,y+1) + U_n(2,0,y-1)$. Combining next Relation (15) for $U_n$ and Inequality (24) proves Relation (14) for $V_n$, for $y \geq 0$. Finally in all cases, Relation (14) is true for $V_n$.

- For Relation (15), we have

$$V_n(1,0,y+1) + V_n(2,0,y+1) \leq U_n(1,0,y+1) + U_n(2,0,y+1) \text{ for } y \geq 0, \tag{25}$$

$$V_n(1,0,y+1) + V_n(2,0,y+1) \leq U_n(2,0,y) + U_n(2,0,y+1) \text{ for } y \geq 0. \tag{26}$$

If "keep" is the optimal action in states $(1,0,y+2)$, for $y \geq 0$, then $V_n(1,0,y+2) + V_n(2,0,y) = U_n(1,0,y+2) + U_n(2,0,y)$. Combining next Relation (15) for $U_n$ and Inequality (25) proves Relation (15) for $V_n$, for $y \geq 0$. If "serve" is the optimal action in state $(1,0,y+2)$, for $y \geq 0$, then $V_n(1,0,y+2) + V_n(2,0,y) = U_n(2,0,y+1) + U_n(2,0,y)$. Inequality (26) proves Relation (15) for $V_n$, for $y \geq 0$. Finally in all cases, Relation (15) is true for $V_n$.

**Step 3.** The step is easy to prove and directly follows from Koole (2007) page 33.

**Step 4.** Assume that $V_n \in \mathcal{F}$, for a given $n \geq 0$. We now prove that $A_n \in \mathcal{F}$. In Relations (5), (7), (8), (9), (11) and (14), $x$ is constant and the arrival of a new call has the same effect on each term of the relation (either increasing the number of customers in queue 1 by one, or changing $z$ into $z+1$). Moreover, since the transition rates are constant, the induction from $V_n$ to $A_n$ is straightforward (see Koole (2007) page 35).

Next, the other relations have to be shown to prove the induction from $V_n$ to $A_n$. For Relations (4), (7), (10) and (13), the case $x < k-1$ is a simplification of the case $k \leq x < N-1$ because the possibility of going to queue 2 is not considered. We therefore only show the case $k \leq x < N-1$.

- For Relation (4), if $x = k-1$, then

$$\begin{aligned}A_n(2,x+1,y) - A_n(2,x,y) &= q_k V_n(2,x+1,y+1) + \alpha_k(V_n(2,x+1,y)+\gamma_3) + (1-\alpha_k-q_k)V_n(2,x+2,y)\\ &- (1-\alpha_{k-1})V_n(2,x+1,y) - \alpha_{k-1}(V_n(2,x,y)+\gamma_3)\\ &= q_k\,(V_n(2,x+1,y+1) - V_n(2,x+1,y)) + \alpha_{k-1}(V_n(2,x+1,y) - V_n(2,x,y))\\ &+ (1-\alpha_k-q_k)(V_n(2,x+2,y) - V_n(2,x+1,y)) + \gamma_3(\alpha_k - \alpha_{k-1}) \geq 0,\end{aligned}$$

since $V_n$ is increasing in $x$ and in $y$.
If $k \leq x < N-1$, then

$$\begin{aligned}A_n(2,x+1,y) - A_n(2,x,y) &= q_{x+1}V_n(2,x+1,y+1) + \alpha_{x+1}(V_n(2,x+1,y)+\gamma_3) + (1-\alpha_{x+1}-q_{x+1})V_n(2,x+2,y)\\ &- q_x V_n(2,x,y+1) - \alpha_x(V_n(2,x,y)+\gamma_3) - (1-\alpha_x-q_x)V_n(2,x+1,y)\\ &= q_x\,(V_n(2,x+1,y+1) - V_n(2,x+1,y)) + (q_{x+1}-q_x)V_n(2,x+1,y+1)\\ &+ \alpha_x(V_n(2,x+1,y) - V_n(2,x,y)) + (\alpha_{x+1}-\alpha_x)V_n(2,x+1,y) + \gamma_3(\alpha_{x+1}-\alpha_x)\\ &+ (1-\alpha_{x+1}-q_{x+1})(V_n(2,x+2,y) - V_n(2,x+1,y)) + (\alpha_x+q_x-\alpha_{x+1}-q_{x+1})V_n(2,x+1,y)\\ &\geq (q_{x+1}-q_x)(V_n(2,x+1,y+1) - V_n(2,x+1,y)) \geq 0,\end{aligned}$$

since $V_n$ is increasing in $y$ and $q_x$ is increasing in $x$.
If $x = N-1$, then

$$\begin{aligned}A_n(2,x+1,y) - A_n(2,x,y) &= q_x V_n(2,x+1,y+1) + (1-q_x)(V_n(2,x+1,y)+\gamma_3)\\ &- q_x V_n(2,x,y+1) - \alpha_x(V_n(2,x,y)+\gamma_3) - (1-\alpha_x-q_x)V_n(2,x+1,y)\\ &= \alpha_x(V_n(2,x+1,y+1) - V_n(2,x,y)) + q_x(V_n(2,x+1,y+1) - V_n(2,x,y+1))\\ &+ \gamma_3(1-q_x-\alpha_x) \geq 0,\end{aligned}$$

since $V_n$ is increasing in $x$ and in $y$. Finally in all cases, Relation (4) is true for $A_n$.

- For Relation (6), we may write

$$\begin{aligned}A_n(2,0,y) - A_n(1,0,y) &= (1-\alpha_0)V_n(2,1,y) + \alpha_0(V_n(2,0,y)+\gamma_3) - V_n(2,0,y)\\ &= (1-\alpha_0)(V_n(2,1,y) - V_n(2,0,y)) + \alpha_0\gamma_3 \geq 0,\end{aligned}$$

since Relation (4) is true for $V_n$. Hence, Relation (6) is true for $A_n$.

For the following relations, we do not write the terms in $\gamma_3$ since the do disappear in the considered differences.

- For Relation (10), if $x = k-1$, then

$$\begin{aligned}&A_n(2,x,y) + A_n(2,x+1,y+1) - A_n(2,x,y+1) - A_n(2,x+1,y)\\ &= \alpha_{k-1}V_n(2,x,y) + (1-\alpha_{k-1})V_n(2,x+1,y) + q_k V_n(2,x+1,y+2) + \alpha_k V_n(2,x+1,y+1) + (1-\alpha_k-q_k)V_n(2,x+2,y+1)\\ &- \alpha_{k-1}V_n(2,x,y+1) - (1-\alpha_{k-1})V_n(2,x+1,y+1) - q_k V_n(2,x+1,y+1) - \alpha_k V_n(2,x+1,y) - (1-\alpha_k-q_k)V_n(2,x+2,y)\\ &= \alpha_{k-1}(V_n(2,x+1,y+1) + V_n(2,x,y) - V_n(2,x+1,y) - V_n(2,x,y+1))\\ &+ q_k(V_n(2,x+1,y+2) + V_n(2,x+2,y) - V_n(2,x+2,y+1) - V_n(2,x+1,y+1))\\ &+ (1-\alpha_k)(V_n(2,x+2,y+1) + V_n(2,x+1,y) - V_n(2,x+1,y+1) - V_n(2,x+2,y)).\end{aligned}$$

The term proportional to $\alpha_{k-1}$ is positive since Relation (10) holds for $V_n$, the term proportional to $q_k$ is positive since Relation (13) holds for $V_n$, the term proportional to $1-\alpha_k$ is positive since Relation (10) holds for $V_n$. Hence, Relation (10) is true for $A_n$, for $x = k-1$.

If $k \leq x < N-1$, then

$$\begin{aligned}
& A_n(2,x,y) + A_n(2,x+1,y+1) - A_n(2,x,y+1) - A_n(2,x+1,y) \\
& = q_x V_n(2,x,y+1) + \alpha_x V_n(2,x,y) + (1-q_x-\alpha_x) V_n(2,x+1,y) \\
& + q_{x+1} V_n(2,x+1,y+2) + \alpha_{x+1} V_n(2,x+1,y+1) + (1-q_{x+1}-\alpha_{x+1}) V_n(2,x+2,y+1) \\
& - q_x V_n(2,x,y+2) - \alpha_x V_n(2,x,y+1) - (1-q_x-\alpha_x) V_n(2,x+1,y+1) \\
& - q_{x+1} V_n(2,x+1,y+1) - \alpha_{x+1} V_n(2,x+1,y) - (1-q_{x+1}-\alpha_{x+1}) V_n(2,x+2,y) \\
& = q_x (V_n(2,x,y+1) + V_n(2,x+1,y+2) - V_n(2,x,y+2) - V_n(2,x+1,y+1)) \\
& + \alpha_x (V_n(2,x,y) + V_n(2,x+1,y+1) - V_n(2,x,y+1) - V_n(2,x+1,y)) \\
& + (1-\alpha_{x+1}-q_{x+1})(V_n(2,x+1,y) + V_n(2,x+2,y+1) - V_n(2,x+1,y+1) - V_n(2,x+2,y)) \\
& + (q_{x+1}-q_x)(V_n(2,x+1,y+2) + V_n(2,x+1,y) - 2V_n(2,x+1,y+1)).
\end{aligned}$$

The terms proportional to $q_x$, $\alpha_x$ and $1-q_{x+1}-\alpha_{x+1}$ are positive since Relation (10) is true for $V_n$, the term proportional to $q_{x+1}-q_x$ is also positive since $V_n$ is convex in $y$. Hence Relation (10) is true for $A_n$, for $k \leq x < N-1$.

If $x = N-1$, then

$$\begin{aligned}
& A_n(2,x,y) + A_n(2,x+1,y+1) - A_n(2,x,y+1) - A_n(2,x+1,y) \\
& = q_x V_n(2,x,y+1) + \alpha_x V_n(2,x,y) + (1-q_x-\alpha_x) V_n(2,x+1,y) + q_x V_n(2,x+1,y+2) + (1-q_x) V_n(2,x+1,y+1) \\
& - q_x V_n(2,x,y+2) - \alpha_x V_n(2,x,y+1) - (1-q_x-\alpha_x) V_n(2,x+1,y+1) - q_x V_n(2,x+1,y+1) - (1-q_x) V_n(2,x+1,y) \\
& = q_x (V_n(2,x,y+1) + V_n(2,x+1,y+2) - V_n(2,x,y+2) - V_n(2,x+1,y+1)) \\
& + \alpha_x (V_n(2,x,y) + V_n(2,x+1,y+1) - V_n(2,x,y+1) - V_n(2,x+1,y)).
\end{aligned}$$

The terms proportional to $q_x$ and $\alpha_x$ are positive since Relation (10) is true for $V_n$. Hence Relation (10) is true for $A_n$, for $x = N-1$.

- For Relation (12), we have for $y \geq 0$,

$$\begin{aligned}
& A_n(1,0,y) + A_n(2,0,y+1) - A_n(2,0,y) - A_n(1,0,y+1) \\
& = V_n(2,0,y) + \alpha_0 V_n(2,0,y+1) + (1-\alpha_0) V_n(2,1,y+1) - \alpha_0 V_n(2,0,y) - (1-\alpha_0) V_n(2,1,y) - V_n(2,0,y+1) \\
& = (1-\alpha_0)(V_n(2,0,y) + V_n(2,1,y+1) - V_n(2,1,y) - V_n(2,0,y+1)) \geq 0,
\end{aligned}$$

since Relation (10) holds for $V_n$. Hence Relation (12) is true for $A_n$.

- For Relation (13), if $x < k-1$ the transition rates are constant and the induction from $V_n$ to $A_n$ follows.

If $x = k - 1$, then

$$\begin{aligned}
&A_n(2,x,y+2) + A_n(2,x+1,y) - A_n(2,x,y+1) - A_n(2,x+1,y+1)\\
&= \alpha_{k-1}V_n(2,x,y+2) + (1-\alpha_{k-1})V_n(2,x+1,y+2) + q_kV_n(2,x+1,y+1) + \alpha_kV_n(2,x+1,y) + (1-\alpha_k-q_k)V_n(2,x+2,y)\\
&- \alpha_{k-1}V_n(2,x,y+1) - (1-\alpha_{k-1})V_n(2,x+1,y+1) - q_kV_n(2,x+1,y+2) - \alpha_kV_n(2,x+1,y+1)\\
&- (1-\alpha_k-q_k)V_n(2,x+2,y+1)\\
&= \alpha_{k-1}(V_n(2,x,y+2) + V_n(2,x+1,y+1) - V_n(2,x+1,y+2) - V_n(2,x,y+1))\\
&+ \alpha_k(V_n(2,x+1,y) + V_n(2,x+2,y+1) - V_n(2,x+2,y) - V_n(2,x+1,y+1))\\
&+ q_k(V_n(2,x+1,y+1) + V_n(2,x+2,y+1) - V_n(2,x+2,y) - V_n(2,x+1,y+2))\\
&+ V_n(2,x+2,y) + V_n(2,x+1,y+2) - V_n(2,x+1,y+1) - V_n(2,x+2,y+1)\\
&= (\alpha_k - \alpha_{k-1})(V_n(2,x+1,y) + V_n(2,x+1,y+2) - 2V_n(2,x+1,y+1))\\
&+ \alpha_{k-1}(V_n(2,x,y+2) + V_n(2,x+1,y) - V_n(2,x,y+1) - V_n(2,x+1,y+1))\\
&(1-q_k-\alpha_k)(V_n(2,x+2,y) + V_n(2,x+1,y+2) - +V_n(2,x+1,y+1) - V_n(2,x+2,y+1)).
\end{aligned}$$

The term proportional to $\alpha_k - \alpha_{k-1}$ is positive since $V_n$ is convex in $y$, the term proportional to $\alpha_{k-1}$ is positive since Relation (13) is true for $V_n$, the term proportional to $1 - q_k - \alpha_k$ is positive since Relation (13) is true for $V_n$. Hence Relation (13) is true for $A_n$, for $x = k - 1$.

If $k \leq x < N - 1$, then

$$\begin{aligned}
&A_n(2,x,y+2) + A_n(2,x+1,y) - A_n(2,x,y+1) - A_n(2,x+1,y+1)\\
&= q_xV_n(2,x,y+3) + \alpha_xV_n(2,x,y+2) + (1-q_x-\alpha_x)V_n(2,x+1,y+2)\\
&+ q_{x+1}V_n(2,x+1,y+1) + \alpha_{x+1}V_n(2,x+1,y) + (1-q_{x+1}-\alpha_{x+1})V_n(2,x+2,y)\\
&- q_xV_n(2,x,y+2) - \alpha_xV_n(2,x,y+1) - (1-q_x-\alpha_x)V_n(2,x+1,y+1)\\
&- q_{x+1}V_n(2,x+1,y+2) - \alpha_{x+1}V_n(2,x+1,y+1) - (1-q_{x+1}-\alpha_{x+1})V_n(2,x+2,y+1)\\
&= q_x(V_n(2,x,y+3) + V_n(2,x+1,y+1) - V_n(2,x+1,y+2) - V_n(2,x,y+2))\\
&+ \alpha_x(V_n(2,x,y+2) + V_n(2,x+1,y) - V_n(2,x+1,y+1) - V_n(2,x,y+1))\\
&+ (1-\alpha_{x+1}-q_{x+1})(V_n(2,x+1,y+2) + V_n(2,x+2,y) - V_n(2,x+1,y+1) - V_n(2,x+2,y+1))\\
&+ (\alpha_{x+1}-\alpha_x)(V_n(2,x+1,y+2) + V_n(2,x+1,y) - 2V_n(2,x+1,y+1)).
\end{aligned}$$

The terms proportional to $q_x$, $\alpha_x$ and $1 - q_{x+1} - \alpha_{x+1}$ are positive since Relation (13) is true for $V_n$, the term proportional to $\alpha_{x+1} - \alpha_x$ is also positive since $V_n$ is convex in $y$. Hence Relation (13) is true for $A_n$, for $k \leq x < N - 1$.

If $x = N - 1$, then

$$\begin{aligned}
&A_n(2,x,y+2) + A_n(2,x+1,y) - A_n(2,x,y+1) - A_n(2,x+1,y+1)\\
&= q_xV_n(2,x,y+3) + \alpha_xV_n(2,x,y+2) + (1-q_x-\alpha_x)V_n(2,x+1,y+2) + q_xV_n(2,x+1,y+1) + (1-q_x)V_n(2,x+1,y)\\
&- q_xV_n(2,x,y+2) - \alpha_xV_n(2,x,y+1) - (1-q_x-\alpha_x)V_n(2,x+1,y+1) - q_xV_n(2,x+1,y+2) - (1-q_x)V_n(2,x+1,y+1)\\
&= q_x(V_n(2,x,y+3) + V_n(2,x+1,y+1) - V_n(2,x+1,y+2) - V_n(2,x,y+2))\\
&+ \alpha_x(V_n(2,x,y+2) + V_n(2,x+1,y) - V_n(2,x,y+1) - V_n(2,x+1,y+1))\\
&+ (1-q_x-\alpha_x)(V_n(2,x+1,y+2) + V_n(2,x+1,y) - 2V_n(2,x+1,y+1)).
\end{aligned}$$

The terms proportional to $q_x$ and $\alpha_x$ are positive since Relation (13) is true for $V_n$, the term proportional to $1 - q_x - \alpha_x$ is also positive since $V_n$ is convex in $y$. Hence Relation (13) is true for $A_n$, for $x = N - 1$.

- For Relation (15), we have

$$\begin{aligned}
&A_n(1,0,y+2) + A_n(2,0,y) - A_n(1,0,y+1) - A_n(2,0,y+1)\\
&= V_n(2,0,y+2) + \alpha_0 V_n(2,0,y) + (1-\alpha_0)V_n(2,1,y) - V_n(2,0,y+1) - \alpha_0 V_n(2,0,y+1) - (1-\alpha_0)V_n(2,1,y+1)\\
&= V_n(2,0,y+2) + V_n(2,1,y) - V_n(2,0,y+1) - V_n(2,1,y+1) + \alpha_0(V_n(2,0,y) + V_n(2,1,y+1) - V_n(2,0,y+1) - V_n(2,1,y)).
\end{aligned}$$

The terms proportional to 1 is positive since Relation (13) is true for $V_n$, the term proportional to $\alpha_0$ is also positive since Relation (10) is true for $V_n$. Hence Relation (15) is true for $A_n$.

**Step 5.** Assume that $V_n \in \mathcal{F}$, for a given $n \geq 0$. We now show that $D_n \in \mathcal{F}$.

- For Relation (4), if $x = 0$, then

$$\begin{aligned}
D_n(2,1,y) - D_n(2,0,y) &= \beta V_n(2,0,y) + \beta\gamma_3 + 2\mu(V_n(2,0,y) - V_n(1,0,y))\\
&+ (1-\lambda-\beta-2\mu)(V_n(2,1,y) - V_n(2,0,y)) - \beta V_n(2,0,y) \geq 0,
\end{aligned}$$

since $V_n$ is increasing in $x$ and Relation (6) is true for $V_n$.

If $x > 0$, then

$$\begin{aligned}
&D_n(2,x+1,y) - D_n(2,x,y) = \beta x(V_n(2,x,y) - V_n(2,x-1,y)) + \beta\gamma_3 + \beta V_n(2,x,y)\\
&+ 2\mu(V_n(2,x,y) - V_n(2,x-1,y)) + (1-\lambda-\beta(x+1)-2\mu)(V_n(2,x+1,y) - V_n(2,x,y)) - \beta V_n(2,x,y) \geq 0,
\end{aligned}$$

since $V_n$ is increasing in $x$. Hence Relation (4) is true for $D_n$.

- For Relation (5), we have

$$D_n(1,0,y) - D_n(0,0,y) = \mu V_n(0,0,y) + (1-\lambda-\mu)(V_n(1,0,y) - V_n(0,0,y)) - \mu V_n(0,0,y) \geq 0.$$

Hence Relation (5) is true for $D_n$.

- For Relation (6), we have

$$D_n(2,0,y) - D_n(1,0,y) = \mu(V_n(1,0,y) - V_n(0,0,y)) + \mu V_n(1,0,y) + (1-\lambda-2\mu)(V_n(2,0,y) - V_n(1,0,y)) - \mu V_n(1,0,y) \geq 0.$$

Hence Relation (6) is true for $D_n$.

- For Relation (7), if $x \geq 0$, then

$$\begin{aligned}
&D_n(2,x,y+1) - D_n(2,x,y) = \beta x(V_n(2,x-1,y+1) - V_n(2,x-1,y)) + 2\mu\mathbf{1}_{(x=0)}(V_n(0,0,y+1) - V_n(0,0,y))\\
&+ (1-\lambda-\beta x-2\mu)(V_n(2,x,y+1) - V_n(2,x,y)) \geq 0,
\end{aligned}$$

since $V_n$ is increasing in $y$. Hence Relation (7) holds for $D_n$.

- Relations (8) and (9) are obviously also true for $D_n$.
- For Relation (10), if $x, y \geq 0$, then

$$\begin{aligned}
&D_n(2,x,y) + D_n(2,x+1,y+1) - D_n(2,x+1,y) - D_n(2,x,y+1)\\
&= \beta x(V_n(2,x-1,y) + V_n(2,x,y+1) - V_n(2,x,y) - V_n(2,x-1,y+1)) + \beta(V_n(2,x,y+1) - V_n(2,x,y))\\
&+ 2\mu\mathbf{1}_{(x=0)}(V_n(1,0,y) + V_n(2,0,y+1) - V_n(2,0,y) - V_n(1,0,y+1))\\
&+ 2\mu\mathbf{1}_{(x>0)}(V_n(2,x-1,y) + V_n(2,x,y+1) - V_n(2,x,y) - V_n(2,x-1,y+1))\\
&+ (1-\lambda-\beta(x+1)-2\mu)(V_n(2,x,y) + V_n(2,x+1,y+1) - V_n(2,x+1,y) - V_n(2,x,y+1))\\
&+ \beta(V_n(2,x,y) - V_n(2,x,y+1)) \geq 0,
\end{aligned}$$

since Relations (10) and (12) are true for $V_n$.

- For Relation (11), we have for $y \geq 0$,

$$\begin{aligned}
&D_n(0,0,y)+D_n(1,0,y+1)-D_n(1,0,y)-D_n(0,0,y+1)\\
&=\mu(V_n(0,0,y+1)-V_n(0,0,y))+(1-\lambda-\mu)(V_n(0,0,y)+V_n(1,0,y+1)-V_n(1,0,y)-V_n(0,0,y+1))\\
&+\mu(V_n(0,0,y)-V_n(0,0,y+1))\geq 0,
\end{aligned}$$

since Relation (11) is true for $V_n$.

- For Relation (12), we have for $y \geq 0$,

$$\begin{aligned}
&D_n(1,0,y)+D_n(2,0,y+1)-D_n(2,0,y)-D_n(1,0,y+1)\\
&=\mu(V_n(0,0,y)-V_n(0,0,y+1))+2\mu(V_n(1,0,y+1)-V_n(1,0,y))\\
&+(1-\lambda-2\mu)(V_n(1,0,y)+V_n(2,0,y+1)-V_n(2,0,y)-V_n(1,0,y+1))+\mu(V_n(1,0,y)-V_n(1,0,y+1))\\
&\geq\mu(V_n(0,0,y)+V_n(1,0,y+1)-V_n(1,0,y)-V_n(0,0,y+1)).
\end{aligned}$$

The term proportional to $\mu$ is positive since Relation (11) is true for $V_n$. Therefore Relation (12) is true for $D_n$.

- For Relation (13), if $x, y \geq 0$, then

$$\begin{aligned}
&D_n(2,x,y+2)+D_n(2,x+1,y)-D_n(2,x,y+1)-D_n(2,x+1,y+1)\\
&=\beta x(V_n(2,x-1,y+2)+V_n(2,x,y)-V_n(2,x-1,y+1)-V_n(2,x,y+1))+\beta(V_n(2,x,y)-V_n(2,x,y+1))\\
&+2\mu\mathbf{1}_{(x=0)}(V_n(1,0,y+2)+V_n(2,0,y)-V_n(1,0,y+1)-V_n(2,0,y+1))\\
&+2\mu\mathbf{1}_{(x>0)}(V_n(2,x-1,y+2)+V_n(2,x,y)-V_n(2,x-1,y+1)-V_n(2,x,y+1))\\
&+(1-\lambda-\beta(x+1)-2\mu)(V_n(2,x,y+2)+V_n(2,x+1,y)-V_n(2,x,y+1)-V_n(2,x+1,y+1))\\
&+\beta(V_n(2,x,y+2)-V_n(2,x,y+1))\geq\beta(V_n(2,x,y+2)+V_n(2,x,y)-2V_n(2,x,y+1))\geq 0,
\end{aligned}$$

since Relations (13) and (15) are true for $V_n$ and $V_n$ is convex in $y$. Therefore, Relation (13) is true for $D_n$.

- For Relation (14), we have for $y \geq 0$,

$$\begin{aligned}
&D_n(0,0,y+2)+D_n(1,0,y)-D_n(0,0,y+1)-D_n(1,0,y+1)\\
&=\mu(V_n(0,0,y)-V_n(0,0,y+1))+(1-\lambda-\mu)(V_n(0,0,y+2)+V_n(1,0,y)-V_n(0,0,y+1)-V_n(1,0,y+1))\\
&+\mu(V_n(0,0,y+2)-V_n(0,0,y+1))\geq\mu(V_n(0,0,y+2)+V_n(0,0,y)-2V_n(0,0,y+1))\geq 0,
\end{aligned}$$

since Relation (14) is true for $V_n$ and since $V_n$ is convex in $y$. Hence, Relation Relation (14) is true for $D_n$.

- For Relation (15), we have for $y \geq 0$,

$$\begin{aligned}
&D_n(1,0,y+2)+D_n(2,0,y)-D_n(1,0,y+1)-D_n(2,0,y+1)\\
&=\mu(V_n(0,0,y+2)-V_n(0,0,y+1))+2\mu(V_n(1,0,y)-V_n(1,0,y+1))\\
&+(1-\lambda-2\mu)(V_n(1,0,y+2)+V_n(2,0,y)-V_n(1,0,y+1)-V_n(2,0,y+1))+\mu(V_n(1,0,y+2)-V_n(1,0,y+1))\\
&\geq\mu(V_n(1,0,y+2)+V_n(1,0,y)-2V_n(1,0,y+1))+\mu(V_n(1,0,y)+V_n(0,0,y+2)-V_n(0,0,y+1)-V_n(1,0,y+1)).
\end{aligned}$$

The two terms proportional to $\mu$ are positive, the first one because $V_n$ is convex in $y$ and the second one because Relation (14) holds for $V_n$. Therefore Relation (15) is true for $D_n$. The proof is completed. □

# B Proof of Proposition 2

To prove Proposition 2, we need to prove by induction on $n$ ($n \geq 0$) that, for $x, y \geq 0$,

$$V'_n(x,y) + V_n(x+1,y) \geq V_n(x,y) + V'_n(x+1,y), \tag{27}$$

$$U'_n(x,y) + U_n(x+1,y) \geq U_n(x,y) + U'_n(x+1,y), \tag{28}$$

$$V'_n(x,y+1) + V_n(x,y) \geq V'_n(x,y+1) + V_n(x,y), \tag{29}$$

$$U'_n(x,y+1) + U_n(x,y) \geq U'_n(x,y+1) + U_n(x,y), \tag{30}$$

where $V_n(x,y)$, $U_n(x,y)$ and $V'_n(x,y)$, $U'_n(x,y)$ are the value functions associated with the parameters $\gamma_1$, $\gamma_2$, $\gamma_3$ and $q$ for $x \geq s+k$, and the parameters $\gamma'_1$, $\gamma'_2$, $\gamma'_3$ and $q+q'$ for $x \geq s+k$, respectively. Summing up Relations (27) and (29) prove that $V'_n(x,y+1) + V_n(x+1,y) \geq V_n(x,y+1) + V'_n(x+1,y)$. This implies that situation 1 requires more reservation than situation 2.

We have $U_0 = V_0 = U'_0 = V'_0 = 0$. Thus, Relations (27), (28), (29) and (30) hold for $n = 0$.

We first prove that Relation (28) implies Relation (27). Assume now that Relation (28) holds for a given $n \geq 0$. Therefore, $U'_n(x,y) + U_n(x+1,y) \geq U'_n(x+1,y) + U_n(x,y)$. We only consider the non-trivial cases where $0 \leq x < s$ and $y > 0$. We have

$$V'_n(x+1,y) + V_n(x,y) \leq U'_n(x+1,y) + U_n(x,y) \text{ for } 0 \leq x \leq s-1, y > 0, \tag{31}$$

$$V'_n(x+1,y) + V_n(x,y) \leq U'_n(x+1,y) + U_n(x+1,y-1) \text{ for } 0 \leq x \leq s-1, y > 0, \tag{32}$$

$$V'_n(x+1,y) + V_n(x,y) \leq U'_n(x+2,y-1) + U_n(x+1,y-1) \text{ for } 0 \leq x \leq s-2, y > 0. \tag{33}$$

If "keep" is the optimal action in states $(x,y)$ and $(x+1,y)$ for situations 2 and 1, respectively, then $V'_n(x,y) + V_n(x+1,y) = U'_n(x,y) + U_n(x+1,y)$. Combining Equation (31) with Relation (28) for $U_n$ proves Relation (27) for $V_n$.

If "serve" is the optimal action in states $(x,y)$ and $(x+1,y)$ for situations 2 and 1, respectively, then $V'_n(x,y) + V_n(x+1,y) = U'_n(x+1,y-1) + U_n(x+2,y-1)$. Combining Equation (33) with Relation (28) for $U_n$ proves Relation (27) for $V_n$.

If "serve" is the optimal action in state $(x,y)$ and "keep" is the optimal action in state $(x+1,y)$ for situations 2 and 1, respectively, then $V'_n(x,y) + V_n(x+1,y) = U'_n(x+1,y-1) + U_n(x+1,y)$. Inequality (32) proves Relation (27) for $V_n$.

The case where "keep" would be the optimal action in state $(x,y)$ and "serve" would be the optimal action in state $(x+1,y)$ for situations 2 and 1, respectively, is not considered because it is in contradiction with Relation (28) for $U_n$.

We second prove that Relation (30) implies Relation (29). Assume now that Relation (30) holds for a given

$n \geq 0$. Again, we only consider the non-trivial cases where $0 \leq x < s$ and $y > 0$. We have

$$V'_n(x,y) + V_n(x,y+1) \leq U'_n(x,y) + U_n(x,y+1) \text{ for } 0 \leq x \leq s-1, y > 0, \tag{34}$$

$$V'_n(x,y) + V_n(x,y+1) \leq U'_n(x,y) + U_n(x+1,y) \text{ for } 0 \leq x \leq s-1, y > 0, \tag{35}$$

$$V'_n(x,y) + V_n(x,y+1) \leq U'_n(x+1,y-1) + U_n(x+1,y) \text{ for } 0 \leq x \leq s-1, y > 0. \tag{36}$$

If "keep" is the optimal action in states $(x,y)$ and $(x,y+1)$ for situations 1 and 2, respectively, then $V_n(x,y)+V'_n(x,y+1) = U_n(x,y)+U'_n(x,y+1)$. Combining Equation (34) with Relation (30) for $U_n$ proves Relation (29) for $V_n$.

If "serve" is the optimal action in states $(x,y)$ and $(x,y+1)$ for situations 1 and 2, respectively, then $V_n(x,y) + V'_n(x,y+1) = U_n(x+1,y-1) + U_n(x+1,y)$. Combining Equation (36) with Relation (30) for $U_n$ proves Relation (29) for $V_n$.

If "keep" is the optimal action in state $(x,y)$ and "serve" is the optimal action in state $(x,y+1)$ for situations 1 and 2, respectively, then $V_n(x,y) + V'_n(x,y+1) = U_n(x,y) + U'_n(x+1,y)$. Inequality (35) proves Relation (29) for $V_n$.

The case where "serve" would be the optimal action in state $(x,y)$ and "keep" would be the optimal action in state $(x,y+1)$ for situations 1 and 2, respectively, is not considered because it is in contradiction with Relation (30) for $U_n$.

We now prove that Relations (27) and (29) for $V_n$ imply Relation (28) and (30) for $U_{n+1}$.

The proof of Relation (27) for the departure term can be easily done since the terms are identical in situations 1 and 2 except for the cost parameter related to the abandonment. This implies a positive difference $(\beta(\gamma_3 - \gamma'_3)((x+1-s)^+ - (x-s)^+) \geq 0$ since $\gamma_3 \geq \gamma'_3$. We therefore only focus on the cost and arrival terms. We denote by $G(x,y)$ and $G'(x,y)$ the cost terms in situations 1 and 2, respectively and $A(x,y)$ and $A'(x,y)$ the arrival terms in situations 1 and 2, respectively. We have

$$G'(x,y) + G(x+1,y) - G(x,y) - G'(x+1,y) = (\gamma_1 - \gamma'_1)((x+1-s)^+ - (x-s)^+) \geq 0,$$

since $\gamma_1 \geq \gamma'_1$ and

$$G(x,y) + G'(x,y+1) - G'(x,y) - G(x,y+1) = \gamma'_2 - \gamma_2 \geq 0,$$

since $\gamma'_2 \geq \gamma_2$.

For the arrival term we may write for $x \geq s+k$ (the terms where $x < s+k$ are simplifications of this case and are therefore omitted)

$$\begin{aligned}
&A'_n(x,y) + A_n(x+1,y) - A_n(x,y) - A'_n(x+1,y) \\
&= \alpha_x(V'_n(x,y) + V_n(x+1,y) - V_n(x,y) - V'_n(x+1,y)) \\
&+ q(V'_n(x,y+1) + V_n(x+1,y+1) - V_n(x,y+1) - V'_n(x+1,y+1)) \\
&+ (1-\alpha_{x+1} - q)(V'_n(x+1,y) + V_n(x+2,y) - V_n(x+1,y) - V'_n(x+2,y)) \\
&+ q'(V'_n(x+2,y) - V'_n(x+1,y+1) + V'_n(x,y+1) - V'_n(x+1,y)) \\
&+ (\gamma_3 - \gamma'_3)(\alpha_{x+1} - \alpha_x).
\end{aligned}$$

The terms proportional to $\alpha_x$, $q$ and $1 - \alpha_{x+1} - q$ are positive since Relation (27) is true for $V_n$. The term proportional to $q'$ is positive since this relation defines that the optimal policy in situation 2 is of switch

type. The last term is also positive since $\gamma_3 \geq \gamma_3'$. Therefore, Relation (28) is true for $U_{n+1}$.
For the arrival term we also may write for $x \geq s+k$ (the terms where $x < s+k$ are simplifications of this case and are therefore omitted)

$$\begin{aligned}
&A_n'(x,y+1) + A_n(x,y) - A_n(x,y+1) - A_n'(x,y)\\
&= \alpha_x(V_n'(x,y+1) + V_n(x,y) - V_n(x,y+1) - V_n'(x,y))\\
&+ q(V_n'(x,y+2) + V_n(x,y+1) - V_n(x,y+2) - V_n'(x,y+1))\\
&+ (1-\alpha_x-q)(V_n'(x+1,y+1) + V_n(x+1,y) - V_n(x+1,y+1) - V_n'(x+1,y))\\
&+ q'(V_n'(x,y+2) - V_n'(x+1,y+1) + V_n'(x+1,y) - V_n'(x,y+1)).
\end{aligned}$$

The terms proportional to $\alpha_x$, $q$ and $1-\alpha_x-q$ are positive since Relation (29) is true for $V_n$. The term proportional to $q'$ is positive since this relation defines that the optimal policy in situation 2 is of switch type. Therefore, Relation (30) is true for $U_{n+1}$. This finishes the proof of the proposition. □

# C Performance Analysis for Model C

We provide here the details for the steps of the performance evaluation method for Model C.

**Step 1.** The stationary probabilities are determined by the following set of equilibrium equations. For $y = 0$, we may write

$$\lambda p_{x,0} = (x+1)\mu p_{x+1,0}, \text{ for } 0 \leq x < y_0, \tag{37}$$

$$\lambda p_{y_0,0} = (y_0+1)\mu p_{y_0+1,0} + y_0\mu p_{y_0,1}, \text{ for } x = y_0, \tag{38}$$

$$(\lambda + x\mu)p_{x,0} = (x+1)\mu p_{x+1,0} + \lambda p_{x-1,0}, \text{ for } y_0 < x < s, \tag{39}$$

$$((1-\alpha)\lambda + s\mu)p_{s,0} = s\mu p_{s+1,0} + \lambda p_{s-1,0}, \text{ for } x = s, \tag{40}$$

$$(\lambda(1-\alpha) + s\mu)p_{x,0} = s\mu p_{x+1,0} + \lambda(1-\alpha)p_{x-1,0}, \text{ for } s < x \leq s+k, \tag{41}$$

$$(\lambda(1-\alpha) + s\mu)p_{x,0} = s\mu p_{x+1,0} + (1-q-\alpha)\lambda p_{x-1,0}, \text{ for } x > s+k. \tag{42}$$

For $y = y_i - 1$ and $1 \leq i \leq s - y_0$, we have

$$(\lambda + (y_0+i-1)\mu)p_{y_0+i-1,y_i-1} = (y_0+i)\mu p_{y_0+i,y_i-1}, \text{ for } x = y_0+i-1, \tag{43}$$

$$(\lambda + (y_0+i)\mu)p_{y_0+i,y_i-1} = \lambda p_{y_0+i-1,y_i-1} + (y_0+i)\mu p_{y_0+i,y_i} + \min(y_0+i+1,s)\mu p_{y_0+i+1,y_i-1}, \text{ for } x = y_0+i. \tag{44}$$

For $0 < y < y_1 - 1$, $y_i \leq y < y_{i+1} - 1$ and $1 \leq i < s - y_0$, we get

$$(\lambda + (y_0+i)\mu)p_{y_0+i,y} = (y_0+i)\mu p_{y_0+i,y+1} + (y_0+i+1)\mu p_{y_0+i+1,y}, \text{ for } x = y_0+i. \tag{45}$$

For $0 < y \leq y_1 - 1$ and $i = 0$ or $y_i \leq y \leq y_{i+1} - 1$ and $1 \leq i < s - y_0$, we have

$$(\lambda + x\mu)p_{x,y} = (x+1)\mu p_{x+1,y} + \lambda p_{x-1,y}, \text{ for } y_0+i < x < s, \tag{46}$$

$$(\lambda(1-\alpha) + s\mu)p_{s,y} = s\mu p_{s+1,y} + \lambda p_{s-1,y}, \text{ for } x = s, \tag{47}$$

$$(\lambda(1-\alpha) + s\mu)p_{x,y} = s\mu p_{x+1,y} + \lambda(1-\alpha)p_{x-1,y}, \text{ for } s < x < s+k, \tag{48}$$

$$(\lambda(1-\alpha) + s\mu)p_{s+k,y} = s\mu p_{s+k+1,y} + \lambda(1-\alpha)p_{s+k-1,y} + q\lambda p_{s+k,y-1}, \text{ for } x = s+k,$$

$$(\lambda(1-\alpha) + s\mu)p_{x,y} = s\mu p_{x+1,y} + (1-q-\alpha)\lambda p_{x-1,y} + q\lambda p_{x,y-1}, \text{ for } x > s+k. \tag{49}$$

Finally, for $y \geq y_{s-y_0}$, we may write

$$(\lambda(1-\alpha)+s\mu)p_{s,y} = s\mu p_{s,y+1} + s\mu p_{s+1,y}, \text{ for } x = s, \tag{50}$$

$$(\lambda(1-\alpha)+s\mu)p_{x,y} = s\mu p_{x+1,y} + \lambda(1-\alpha)p_{x-1,y}, \text{ for } s < x < s+k,$$

$$(\lambda(1-\alpha)+s\mu)p_{s+k,y} = s\mu p_{s+k+1,y} + \lambda(1-\alpha)p_{s+k-1,y} + q\lambda p_{s+k,y-1}, \text{ for } x = s+k,$$

$$(\lambda(1-\alpha)+s\mu)p_{x,y} = s\mu p_{x+1,y} + (1-q-\alpha)\lambda p_{x-1,y} + q\lambda p_{x,y-1}, \text{ for } x > s+k. \tag{51}$$

**Step 2.** We denote by $a$ the offered load, $a = \frac{\lambda}{\mu}$. Lemma 1 simplifies the expressions of $p_{x,y}$, for $x \leq s+k$ and $y \geq 0$, by writing them as a function of only two state probabilities from the row $y$ in the Markov chain as given in Figure 7.

**Lemma 1** *The following holds.*

1. *If $y = y_i - 1$ for $1 \leq i \leq s - y_0$, then*

$$p_{y_0+i,y_i-1} = \left(\frac{a}{y_0+i} + \frac{y_0+i-1}{y_0+i}\right) p_{y_0+i-1,y_i-1}.$$

2. *For $1 \leq i < s - y_0$, $y_i - 1 \leq y < y_{i+1} - 1$ and $2 \leq x \leq s - y_0 - i$ or $i = 0$, $0 \leq y < y_1 - 1$ and $2 \leq x \leq s - y_0$, we have*

$$p_{y_0+i+x,y} = \frac{1}{(y_0+i+x)!}\left((y_0+i+1)p_{y_0+i+1,y}\sum_{j=0}^{x-1}(y_0+i+j)!a^{x-j-1} - ap_{y_0+i,y}\sum_{j=1}^{x-1}(y_0+i+j)!a^{x-j-1}\right).$$

3. *For $0 \leq y < y_{s-y_0} - 1$, we have*

$$p_{s+1,y} = \left(1 + \frac{a(1-\alpha)}{s}\right) p_{s,y} - \frac{a}{s} p_{s-1,y}.$$

4. *For $y \geq 0$ and $0 \leq x \leq k$, we have*

$$p_{s+x,y} = \left(1 - \frac{a(1-\alpha)}{s}\right)^{-1} \left(p_{s+1,y}\left(1 - \left(\frac{a(1-\alpha)}{s}\right)^x\right) - p_{s,y}\left(\frac{a(1-\alpha)}{s} - \left(\frac{a(1-\alpha)}{s}\right)^x\right)\right).$$

***Proof.*** The proof of the first statement is straightforward. If $y = y_i - 1$ for $1 \leq i \leq s - y_0$, then Equation (43) leads to $p_{y_0+i,y_i-1} = (\frac{a}{y_0+i} + \frac{y_0+i-1}{y_0+i})p_{y_0+i-1,y_i-1}$.

We now prove the second statement by induction on $x$. For $1 \leq i < s - y_0$, $y_i - 1 \leq y < y_{i+1} - 1$ and $2 \leq x \leq s - y_0 - i$ or $i = 0$, $0 \leq y < y_1 - 1$ and $2 \leq x \leq s - y_0$, let us define the property $P(x)$ by

$$P(x) : p_{y_0+i+x,y} = \frac{1}{(y_0+i+x)!}\left((y_0+i+1)p_{y_0+i+1,y}\sum_{j=0}^{x-1}(y_0+i+j)!a^{x-j-1} - ap_{y_0+i,y}\sum_{j=1}^{x-1}(y_0+i+j)!a^{x-j-1}\right),$$

for $0 \leq i < s - y_0$ and $2 \leq x < s - y_0 - i - 1$.

Combining $x = y_0 + i + 1$ and Equation (46), and combining $x = y_0 + 1$ and Equation (39) prove that $P(2)$ is true.

Assume that $P(x)$ and $P(x+1)$ are true, and let us prove that $P(x+2)$ is also true, for $0 \leq i < s - y_0$ and $2 \leq x < s - y_0 - i - 1$.

We may write

$$p_{y_0+i+x,y} = \frac{1}{(y_0+i+x)!}\left((y_0+i+1)p_{y_0+i+1,y}\sum_{j=0}^{x-1}(y_0+i+j)!a^{x-j-1} - ap_{y_0+i,y}\sum_{j=1}^{x-1}(y_0+i+j)!a^{x-j-1}\right),$$

and

$$p_{y_0+i+x+1,y} = \frac{1}{(y_0+i+x+1)!}\left((y_0+i+1)p_{y_0+i+1,y}\sum_{j=0}^{x}(y_0+i+j)!a^{x-j} - ap_{y_0+i,y}\sum_{j=1}^{x}(y_0+i+j)!a^{x-j}\right).$$

Equation (46) for $1 \leq i < s - y_0$ and $2 \leq x < s - y_0 - i - 1$ or Equation (39) for $i = 0$ and $2 \leq x < s - y_0 - 1$ are equivalent to

$$p_{y_0+i+x+2,y} = \frac{a+y_0+i+x+1}{y_0+i+x+2}p_{y_0+i+x+1,y} - \frac{a}{y_0+i+x+2}p_{y_0+i+x,y}.$$

We thus obtain, for $0 \leq i < s - y_0$ and $2 \leq x < s - y_0 - i - 1$,

$$\begin{aligned}
&p_{y_0+i+x+2,y} = \\
&\frac{a+y_0+i+x+1}{y_0+i+x+2}\left(\frac{1}{(y_0+i+x+1)!}\left((y_0+i+1)p_{y_0+i+1,y}\sum_{j=0}^{x}(y_0+i+j)!a^{x-j} - ap_{y_0+i,y}\sum_{j=1}^{x}(y_0+i+j)!a^{x-j}\right)\right)\\
&-\frac{a}{y_0+i+x+2}\left(\frac{1}{(y_0+i+x)!}\left((y_0+i+1)p_{y_0+i+1,y}\sum_{j=0}^{x-1}(y_0+i+j)!a^{x-j-1} - ap_{y_0+i,y}\sum_{j=1}^{x-1}(y_0+i+j)!a^{x-j-1}\right)\right)\\
&= \frac{(y_0+i+1)p_{y_0+i+1,y}}{(y_0+i+x+2)!}\left((a+y_0+i+x+1)\sum_{j=0}^{x}(y_0+i+j)!a^{x-j} - a(y_0+i+x+1)\sum_{j=0}^{x-1}(y_0+i+j)!a^{x-j-1}\right)\\
&-\frac{ap_{y_0+i,y}}{(y_0+i+x+2)!}\left((a+y_0+i+x+1)\sum_{j=1}^{x}(y_0+i+j)!a^{x-j} - a(y_0+i+x+1)\sum_{j=1}^{x-1}(y_0+i+j)!a^{x-j-1}\right)\\
&= \frac{1}{(y_0+i+x+2)!}\left((y_0+i+1)p_{y_0+i+1,y}\sum_{j=0}^{x+1}(y_0+i+j)!a^{x-j+1} - ap_{y_0+i,y}\sum_{j=1}^{x+1}(y_0+i+j)!a^{x-j+1}\right).
\end{aligned}$$

We next deduce that $P(x+2)$ is also true for $0 \leq i < s - y_0$ and $2 \leq x < s - y_0 - i - 1$. So, the property $P(x)$ is true, which finishes the proof of the second statement.

The third statement immediately follows from Equations (40) and (47).

Let us now prove the fourth statement. The corresponding homogeneous equation to Equations (41) and (48) is

$$s\mu z^2 - (\lambda(1-\alpha) + s\mu)z + \lambda(1-\alpha) = 0,$$

with $z$ as a variable, for $z \in \mathbb{C}$. It has two solutions, $z = 1$ and $z = \frac{a(1-\alpha)}{s}$. Thus for $y \geq 0$ and $0 \leq x \leq k$, $p_{s+x,y} = \alpha + \beta\left(\frac{a(1-\alpha)}{s}\right)^x$ with $p_{s,y} = \alpha + \beta$ and $p_{s+1,y} = \alpha + \beta\frac{a(1-\alpha)}{s}$. Finally, for $y \geq 0$ and $0 \leq x \leq k$, we obtain

$$p_{s+x,y} = \left(1 - \frac{a(1-\alpha)}{s}\right)^{-1}\left(p_{s+1,y}\left(1 - \left(\frac{a(1-\alpha)}{s}\right)^x\right) - p_{s,y}\left(\frac{a(1-\alpha)}{s} - \left(\frac{a(1-\alpha)}{s}\right)^x\right)\right).$$

This finishes the proof of the fourth statement, and that of the lemma. □

**Step 3.** We show in Lemma 2 how $p_{x,y}$, for $x \geq s+k$ and $y \geq 0$, can be computed as a function of $p_{s+k,0}, p_{s+k,1}, \cdots, p_{s+k,y}$.

**Lemma 2** *The solution of Equations (42), (49) and (51) is given by*

$$p_{x+s+k,y} = \left( \sum_{j=0}^{y} a_{y,j} x^j \right) z^x, \tag{52}$$

*for $x \geq 0$, where*

$$z = \frac{1}{2} \left( 1 + \frac{a(1-\alpha)}{s} - \sqrt{\left(1 + \frac{a(1-\alpha)}{s}\right)^2 - \frac{4(1-q-\alpha)a}{s}} \right),$$

*and the constants $a_{y,j}$ for $y \geq 0$ and $0 \leq j \leq y$ are given by*

$$a_{y,0} = p_{s+k,y}, \tag{53}$$

*for $y \geq 0$,*

$$a_{y,y} = \frac{a_{0,0}}{y!} \left( \frac{q\lambda z}{-s\mu z^2 + (1-q-\alpha)\lambda} \right)^y,$$

*for $y > 0$, and*

$$a_{y,j+1} = \left[ (-s\mu z^2 + (1-q-\alpha)\lambda)(j+1) \right]^{-1} \left( \sum_{i=j+2}^{y} a_{y,i} \binom{i}{j} \left( s\mu z^2 + (-1)^{i+j}(1-q-\alpha)\lambda \right) + q\lambda z a_{y-1,j} \right), \tag{54}$$

*for $0 \leq j < y-1$ and $y > 1$.*

***Proof.*** Consider the system of equations given by Equations (42), (49) and (51). This system can be solved analytically using standard results from the theory of linear difference equations. Consider the corresponding homogeneous equation to Equations (42), (49) and (51). We have

$$s\mu z^2 - (\lambda(1-\alpha) + s\mu)z + (1-q-\alpha)\lambda = 0, \tag{55}$$

with $z$ as a variable, for $z \in \mathbb{C}$. It has two solutions denoted by $z$ and $z'$ and are given by

$$z = \frac{1}{2} \left( 1 + \frac{a(1-\alpha)}{s} - \sqrt{\left(1 + \frac{a(1-\alpha)}{s}\right)^2 - \frac{4(1-q-\alpha)a}{s}} \right),$$

and

$$z' = \frac{1}{2} \left( 1 + \frac{a(1-\alpha)}{s} + \sqrt{\left(1 + \frac{a(1-\alpha)}{s}\right)^2 - \frac{4(1-q-\alpha)a}{s}} \right).$$

We next provide the intervals where $z$ and $z'$ are ranging. We have $0 \leq z < 1$ and $z' \geq 1$. Let us first prove that $z' \geq 1$. Since $z'$ increases in $q$, we have

$$z' \geq \frac{1}{2}\left(1+\frac{a(1-\alpha)}{s}+\sqrt{\left(1+\frac{a(1-\alpha)}{s}\right)^2-\frac{4a(1-\alpha)}{s}}\right)=\frac{1}{2}\left(1+\frac{a(1-\alpha)}{s}+\sqrt{\left(1-\frac{a(1-\alpha)}{s}\right)^2}\right)=1.$$

In what follows, we prove that $0 \leq z < 1$. Since $z$ decreases in $q$,

$$z \leq \frac{1}{2}\left(1+\frac{a(1-\alpha)}{s}-\sqrt{\left(1+\frac{a(1-\alpha)}{s}\right)^2-\frac{4a(1-\alpha)}{s}}\right)=\frac{1}{2}\left(1+\frac{a(1-\alpha)}{s}-\sqrt{\left(1-\frac{a(1-\alpha)}{s}\right)^2}\right)=\frac{a(1-\alpha)}{s}<1.$$

From Equation (55), we may write $s\mu zz' = (1-q)\lambda$. Since $\lambda \geq 0$, $0 \leq q \leq 1$ and $z' > 1 > 0$, we obtain $z \geq 0$.

Because of the last term in the right hand side of Equations (49) and (51), the stationary probabilities $p_{x+s+k,y}$, for $x \geq 0$ and $y \geq 0$, can be written as a sum of two polynomials multiplied by $z^x$ and $z'^x$, respectively. Since $z' > 1$, the convergence of the stationary probabilities forces the polynomial that is multiplied by $z'^x$ to be equal to zero. We therefore obtain Equation (52), for $x \geq 0$ and $y \geq 0$, that is,

$$p_{x+s+k,y} = \left(\sum_{j=0}^{y} a_{y,j}x^j\right) z^x,$$

with $a_{y,j} \in \mathbb{R}$ for $y \geq 0$ and $0 \leq j \leq y$. In what follows, we compute the parameters $a_{y,j}$, for $y \geq 0$ and $0 \leq j \leq y$, as a function of $p_{s+k,y}$, for $y \geq 0$. It is straightforward to obtain Equation (53). Using Equations (49), (51) and (52), we have

$$\begin{aligned}(\lambda(1-\alpha)+s\mu)\left(\sum_{j=0}^{y} a_{y,j}x^j\right) z^x = s\mu\left(\sum_{j=0}^{y} a_{y,j}(x+1)^j\right) z^{x+1} \\ + (1-q-\alpha)\lambda\left(\sum_{j=0}^{y} a_{y,j}(x-1)^j\right) z^{x-1} + q\lambda\left(\sum_{j=0}^{y-1} a_{y-1,j}x^j\right) z^x,\end{aligned} \tag{56}$$

for $x, y > 0$. Since

$$\sum_{j=0}^{y} a_{y,j}(x+1)^j = \sum_{j=0}^{y}\left(\sum_{i=j}^{y} a_{y,i}\binom{i}{j}\right) x^j,$$

and

$$\sum_{j=0}^{y} a_{y,j}(x-1)^j = \sum_{j=0}^{y}\left(\sum_{i=j}^{y} (-1)^i a_{y,i}\binom{i}{j}\right) (-1)^j x^j,$$

Equation (56) leads to

$$(\lambda(1-\alpha)+s\mu)a_{y,y-1}z = s\mu z^2(a_{y,y-1}+a_{y,y}y) + (1-q-\alpha)\lambda(a_{y,y-1}-a_{y,y}y) + q\lambda z a_{y-1,y-1}, \tag{57}$$

for $y > 0$. Since $z$ is a root of Equation (55), Equation (57) can be rewritten as

$$0 = s\mu z^2 a_{y,y}y - (1-q-\alpha)\lambda a_{y,y}y + q\lambda z a_{y-1,y-1},$$

for $y > 0$. This implies

$$a_{y,y} = a_{y-1,y-1}\frac{q\lambda z}{(-s\mu z^2 + (1-q-\alpha)\lambda)y},$$

for $y > 0$. It thus follows that

$$a_{y,y} = \frac{a_{0,0}}{y!}\left(\frac{q\lambda z}{-s\mu z^2 + (1-q-\alpha)\lambda}\right)^y,$$

for $y > 0$. For $0 \le j < y-1$ and $y > 1$, Equation (56) also leads to

$$(\lambda(1-\alpha) + s\mu)a_{y,j}z = s\mu z^2\left(\sum_{i=j}^{y} a_{y,i}\binom{i}{j}\right) + (1-q-\alpha)\lambda\left(\sum_{i=j}^{y}(-1)^i a_{y,i}\binom{i}{j}\right)(-1)^j + q\lambda z a_{y-1,j}. \tag{58}$$

Since $z$ is a root of Equation (55), Equation (58) can be rewritten as

$$0 = a_{y,j+1}(j+1)(s\mu z^2 - (1-q-\alpha)\lambda) + \sum_{i=j+2}^{y} a_{y,i}\binom{i}{j}\left(s\mu z^2 + (-1)^{i+j}(1-q-\alpha)\lambda\right) + q\lambda z a_{y-1,j},$$

for $0 \le j < y-1$ and $y > 1$. Finally, this leads to Equation (54).

We then compute $a_{y,j}$, for $y \ge 0$ and $0 \le j \le y$, as a function of $p_{s+k,y}$, for $y \ge 0$. This finishes the proof of the lemma. □

**Step 4.** Here, we evaluate all stationary probabilities for $x \ge 0$ and $y = 0$ as a function of $p_{0,0}$. Using Equation (37), we have

$$p_{x,0} = \frac{a^x}{x!}p_{0,0}, \tag{59}$$

for $0 \le x \le y_0$. Using the second statement of Lemma 1, we obtain

$$p_{y_0+x,0} = \frac{1}{(y_0+x)!}\left((y_0+1)p_{y_0+1,0}\sum_{j=0}^{x-1}(y_0+j)!a^{x-j-1} - ap_{y_0,0}\sum_{j=1}^{x-1}(y_0+j)!a^{x-j-1}\right), \tag{60}$$

for $2 \le x \le s - y_0$. From the second and third statements of Lemma 1, we may write

$$p_{s+x,0} = \left(1 - \frac{a(1-\alpha)}{s}\right)^{-1}\left(p_{s,0}\left(1 - \left(\frac{a(1-\alpha)}{s}\right)^{x+1}\right) - p_{s-1,0}\frac{a}{s}\left(1 - \left(\frac{a(1-\alpha)}{s}\right)^{x}\right)\right), \tag{61}$$

for $0 \leq x \leq k$. Using now Equations (60) and (61), we obtain

$$p_{s+k,0} = \frac{(y_0+1)p_{y_0+1,0}}{s!(1-\frac{a(1-\alpha)}{s})}\left(\left(\frac{a(1-\alpha)}{s}\right)^k\left(1-\frac{a}{s}\right)\sum_{j=0}^{s-y_0-2}(y_0+j)!a^{s-y_0-j-1}+(s-1)!\left(1-\left(\frac{a(1-\alpha)}{s}\right)^{k+1}\right)\right) \tag{62}$$

$$-\frac{ap_{y_0,0}}{s!(1-\frac{a(1-\alpha)}{s})}\left(\left(\frac{a(1-\alpha)}{s}\right)^k\left(1-\frac{a}{s}\right)\sum_{j=1}^{s-y_0-2}(y_0+j)!a^{s-y_0-j-1}+(s-1)!\left(1-\left(\frac{a(1-\alpha)}{s}\right)^{k+1}\right)\right).$$

Next, combining Equation (38) and $y_0\mu p_{y_0,1} = q\lambda\sum\limits_{x=0}^{\infty}p_{s+k+x,0}$ provides a relation between $p_{y_0,0}$ and $p_{y_0+1,0}$. From Equations (38)-(42), we have

$$\lambda p_{y_0,0} - (y_0+1)\mu p_{y_0+1,0} = q\lambda\frac{p_{s+k,0}}{1-z}.$$

Combining the previous equation and Equations (59) and (62) implies

$$p_{y_0+1,0} = \frac{a}{y_0+1}\frac{1+\frac{qa}{(1-a(1-\alpha)/s)(1-z)s!}\left(\left(\frac{a(1-\alpha)}{s}\right)^k(1-\frac{a}{s})\sum\limits_{j=1}^{s-y_0-2}(y_0+j)!a^{s-y_0-j-1}+(s-1)!\left(1-\left(\frac{a(1-\alpha)}{s}\right)^{k+1}\right)\right)}{1+\frac{qa}{(1-a(1-\alpha)/s)(1-z)s!}\left(\left(\frac{a(1-\alpha)}{s}\right)^k(1-\frac{a}{s})\sum\limits_{j=0}^{s-y_0-2}(y_0+j)!a^{s-y_0-j-1}+(s-1)!\left(1-\left(\frac{a(1-\alpha)}{s}\right)^{k+1}\right)\right)}\frac{a^{y_0}}{y_0!}p_{0,0}. \tag{63}$$

Using Equations (38) and (63), we also obtain

$$p_{y_0,1} = \frac{p_{0,0}}{1-z}\frac{q\frac{a}{y_0}\frac{a^s}{s!}\left(\frac{a(1-\alpha)}{s}\right)^k}{1+\frac{qa}{(1-a(1-\alpha)/s)(1-z)s!}\left(\left(\frac{a(1-\alpha)}{s}\right)^k(1-\frac{a}{s})\sum\limits_{j=0}^{s-y_0-2}(y_0+j)!a^{s-y_0-j-1}+(s-1)!\left(1-\left(\frac{a(1-\alpha)}{s}\right)^{k+1}\right)\right)}.$$

Using Lemmas 1 and 2 together with Equations (63) and (59), we thus have closed-form expressions for the stationary probabilities $p_{x,0}$ for $x \geq 0$ and $p_{y_0,1}$ as function of $p_{0,0}$.

**Step 5.** We propose in this step a method to compute the stationary probabilities of a given row as a function of the stationary probabilities in the previous rows of the Markov chain. Consider $y \geq 0$, and suppose that the stationary probabilities of rows $0, 1, \cdots, y$ are known in the Markov chain as a function of $p_{0,0}$. If for a given $i$ ($i \in \{1, \cdots s-y_0-1\}$) we have $y_i \leq y+1 < y_{i+1}-1$ or $0 < y+1 < y_1-1$, then $(y_0+i)\mu p_{y_0+i,y+1} = q\lambda\sum\limits_{x=0}^{\infty}p_{s+k+x,y}$, and if $y+1 \geq y_{s-y_0}$ then $s\mu p_{s,y+1} = q\lambda\sum\limits_{x=0}^{\infty}p_{s+k+x,y}$. Consequently, the first stationary probability of row $y+1$ is also known as a function of $p_{0,0}$.

Observe that using Equation (52) for $y \geq 0$, we have

$$\sum_{x=0}^{\infty}p_{s+k+x,y} = \sum_{x=0}^{\infty}\left(\sum_{j=0}^{y}a_{y,j}x^j\right)z^x = \sum_{j=0}^{y}a_{y,j}\left(\sum_{x=0}^{\infty}x^jz^x\right).$$

For $0 \leq j \leq y$, and $x, y \geq 0$, we define the function $f_j$ in the variable $t$ by $f_j(t) = \sum\limits_{x=0}^{\infty}x^jt^x$ with $t \in [0,1)$. The function $f_j(t)$ is given by the recursive relation $f_{n+1}(t) = t\,(f_n(t))'$ for $n \geq 0$ and $f_0(t) = \frac{1}{1-t}$ with

$t \in [0,1)$ (Queffélec and Zuily, 2013). Thus we can derive the infinite sum $\sum_{x=0}^{\infty} p_{s+k+x,y}$, for $0 \le j \le y$, and $x, y \ge 0$, through a finite number of calculations. □

We next distinguish three cases.

- *Case 1*: If for a given $i$ ($i \in \{1, \cdots s - y_0\}$) $y+1 = y_i - 1$, then using the first statement of Lemma 1, the second stationary probability of row $y+1$ ($p_{y_0+i,y_i-1}$) is also known as a function of $p_{0,0}$. Using Lemma 1 we evaluate $p_{x,y_i-1}$ for $y_0+i-1 \le x \le s+k$ as a function of $p_{0,0}$ and $p_{y_0+i+1,y_i-1}$. Using Lemma 2 we evaluate $p_{s+k+x,y_i-1}$ for $x \ge 0$ as a function of $p_{s+k,0}, p_{s+k,1}, \cdots, p_{s+k,y_i-1}$. Since the stationary probabilities of rows $0, 1, \cdots, y_i - 2$ are known as a function of $p_{0,0}$ then we evaluate $p_{s+k+x,y_i-1}$ for $x \ge 0$ as a function of $p_{0,0}$ and $p_{y_0+i+1,y_i-1}$. Using Equation (44), we obtain $(y_0+i)\mu p_{y_0+i,y_i} = (\lambda(1-\mathbf{1}_{y_0+i\ge s}\alpha)+(y_0+i)\mu)p_{y_0+i,y_i-1} - \lambda p_{y_0+i-1,y_i-1} - \min(y_0+i+1,s)\mu p_{y_0+i+1,y_i-1}$. Moreover, we have $(y_0+i)\mu p_{y_0+i,y_i} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,y_i-1}$. Thus the equation $(\lambda(1-\mathbf{1}_{y_0+i\ge s}\alpha)+(y_0+i)\mu)p_{y_0+i,y_i-1} - \lambda p_{y_0+i-1,y_i-1} - \min(y_0+i+1,s)\mu p_{y_0+i+1,y_i-1} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,y_i-1}$ provides a relation between $p_{0,0}$ and $p_{y_0+i+1,y_i-1}$. As a consequence all probabilities of row $y+1$ can be derived as a function of $p_{0,0}$.

- *Case 2*: If for a given $i$ ($i \in \{1, \cdots s - y_0 - 1\}$) we have $y_i \le y+1 < y_{i+1} - 1$ or $0 < y+1 < y_1 - 1$, then using Lemma 1 we evaluate $p_{x,y+1}$ for $y_0+i \le x \le s+k$ as a function of $p_{0,0}$ and $p_{y_0+i+1,y+1}$. Using Lemma 2, we evaluate $p_{s+k+x,y+1}$ for $x \ge 0$ as a function of $p_{s+k,0}, p_{s+k,1}, \cdots, p_{s+k,y+1}$. Since the stationary probabilities of rows $0, 1, \cdots, y$ are known as a function of $p_{0,0}$ then we evaluate $p_{s+k+x,y+1}$ for $x \ge 0$ as a function of $p_{0,0}$ and $p_{y_0+i+1,y+1}$. Using Equation (45) we obtain $(y_0+i)\mu p_{y_0+i,y+2} = (\lambda(1-\mathbf{1}_{y_0+i\ge s}\alpha)+(y_0+i)\mu)p_{y_0+i,y+1} - (y_0+i+1)\mu p_{y_0+i+1,y+1}$. Moreover we have $s\mu p_{s,y+2} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,y+1}$. Thus the equation $(\lambda(1-\mathbf{1}_{y_0+i\ge s}\alpha)+(y_0+i)\mu)p_{y_0+i,y+1} - (y_0+i+1)\mu p_{y_0+i+1,y+1} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,y+1}$ provides a relation between $p_{0,0}$ and $p_{y_0+i+1,y+1}$. As a consequence all probabilities of row $y+1$ can be derived as a function of $p_{0,0}$.

- *Case 3*: If $y+1 \ge y_{s-y_0}$, then using Lemma 1 we evaluate $p_{x,y+1}$ for $s \le x \le s+k$ as a function of $p_{0,0}$ and $p_{s+1,y+1}$. In all cases using Lemma 2 we evaluate $p_{s+k+x,y+1}$ for $x \ge 0$ as a function of $p_{s+k,0}, p_{s+k,1}, \cdots, p_{s+k,y+1}$. Since the stationary probabilities of rows $0, 1, \cdots, y$ are known as a function of $p_{0,0}$, then we evaluate $p_{s+k+x,y+1}$ for $x \ge 0$ as a function of $p_{0,0}$ and $p_{s+1,y+1}$. Using Equation (50), we obtain $s\mu p_{s,y+2} = (\lambda(1-\alpha)+s\mu)p_{s,y+1} - s\mu p_{s+1,y+1}$. Moreover, we have $(y_0+i)\mu p_{y_0+i,y_i} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,y_i-1}$. Thus, the equation $(\lambda(1-\alpha)+s\mu)p_{s,y+1} - s\mu p_{s+1,y+1} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,y_i-1}$ provides a relation between $p_{0,0}$ and $p_{s+1,y+1}$. As a consequence all probabilities of row $y+1$ can be derived as a function of $p_{0,0}$.

**Step 6.** We now evaluate $p_{0,0}$. In what follows we prove that the overall sum of the probabilities can be evaluated after a finite number of calculations. We define the quantity $P_x$ as $P_x = \sum_{y=0}^{\infty} p_{x,y}$ for $x \ge s$. For $s \le x < s+k$ we have $\lambda(1-\alpha)P_x = s\mu P_{x+1}$, then $P_{s+x} = \left(\frac{a(1-\alpha)}{s}\right)^x P_s$, for $0 \le x \le k$. For $x \ge s+k$ we have $(1-q-\alpha)\lambda P_x = s\mu P_{x+1}$, then $P_{s+k+x} = \left(\frac{a(1-\alpha)}{s}\right)^k \left(\frac{(1-q-\alpha)a}{s}\right)^x P_s$, for $x \ge 0$. Using now $y_0\mu \sum_{y=1}^{y_1-1} p_{y_0,y} + (y_0+1)\mu \sum_{y=y_1}^{y_2-1} p_{y_0+1,y} + \cdots + (s-1)\mu \sum_{y=y_{s-y_0-1}}^{y_{s-y_0}-1} p_{s-1,y} + s\mu \left(P_s - \sum_{y=0}^{y_{s-y_0}-1} p_{s,y}\right) = \lambda q \sum_{x=0}^{\infty} P_{s+k+x}$, and $\sum\limits_{x=0}^{\infty} P_{s+k+x} = P_s \frac{\left(\frac{a(1-\alpha)}{s}\right)^k}{1-\frac{a(1-q-\alpha)}{s}}$, we therefore obtain

$$P_s = \left[1 - q\frac{a}{s}\left(\frac{a(1-\alpha)}{s}\right)^k \frac{1}{1-\frac{a(1-q-\alpha)}{s}}\right]^{-1} \left(\sum_{y=0}^{y_{s-y_0}-1} p_{s,y} - \frac{y_0}{s}\sum_{y=1}^{y_1-1} p_{y_0,y} - \frac{y_0+1}{s}\sum_{y=y_1}^{y_2-1} p_{y_0+1,y} - \cdots - \frac{s-1}{s}\sum_{y=y_{s-y_0-1}}^{y_{s-y_0}-1} p_{s-1,y}\right).$$

Thus the quantity $P_s$ can be computed after a finite number of calculations as a function of $p_{0,0}$. Since $\sum\limits_{x=s}^{\infty} P_x = P_s \sum\limits_{x=0}^{k-1}\left(\frac{a(1-\alpha)}{s}\right)^x + P_s\frac{(a(1-\alpha))^k}{s^k}\frac{1}{1-\frac{a(1-q-\alpha)}{s}}$, the overall sum of the probabilities can also be evaluated after a finite number of calculations. Using the fact that all probabilities sum up to one, we obtain $p_{0,0}$. This finishes the characterization of all stationary probabilities.

**Step 7.** We now use the stationary probabilities to derive the system performance measures. The proportion of customers who ask for a callback, $\psi$, is given by

$$\psi = q\sum_{x=s+k}^{\infty} P_x = P_s \frac{q\left(\frac{a(1-\alpha)}{s}\right)^k}{1-\frac{a(1-q-\alpha)}{s}}.$$

The proportion of customers who balk the system, $P_b$, is given by

$$P_b = \alpha\sum_{x=s}^{\infty} P_x = \alpha P_s\left(\sum_{x=0}^{k-1}\left(\frac{a(1-\alpha)}{s}\right)^x + \frac{(a(1-\alpha))^k}{s^k}\frac{1}{1-\frac{a(1-q-\alpha)}{s}}\right).$$

Applying the Little law leads to $\lambda(1-\psi-P_b)E(W_1) = \sum\limits_{x=0}^{\infty} xP_{s+x}$. Therefore,

$$E(W_1) = \frac{P_s}{\lambda(1-\psi-P_b)}\left(\left(\frac{a(1-\alpha)}{s}\right)\frac{1-k\left(\frac{a(1-\alpha)}{s}\right)^{k-1}+(k-1)\left(\frac{a(1-\alpha)}{s}\right)^k}{\left(1-\frac{a(1-\alpha)}{s}\right)^2} + \left(\frac{a(1-\alpha)}{s}\right)^k\frac{\frac{(1-q-\alpha)a}{s}+k\left(1-\frac{(1-q-\alpha)a}{s}\right)}{\left(1-\frac{(1-q-\alpha)a}{s}\right)^2}\right).$$

Again, applying the Little law implies

$$E(W_2) = \frac{1}{\lambda\psi}\left(\sum_{y=1}^{y_1-1} yp_{y_0,y} + \sum_{y=1}^{y_2-1} yp_{y_0+1,y} + \cdots + \sum_{y=1}^{y_{s-y_0}-1} yp_{s-1,y} + \sum_{y=1}^{\infty}\sum_{x=s}^{\infty} yp_{x,y}\right).$$

Using now the following relation

$$\psi E(W_2) + (1-\psi-P_b)E(W_1) = E(W),$$

we obtain $E(W)$. This finishes the characterization of the performance measures in the general case.

# D Highest Reservation and Non-Idling Cases

We simplify here the expressions given in Corollary 2. We focus on the multi-server setting for the special cases $y_0 = 1$ (highest reservation) and $y_0 = s$ (non-idling). These results are for example useful for the numerical computations in Table 2. We first consider the highest reservation policy with $y_0 = 1$. We have

$$\Psi = \frac{q\left(\frac{a(1-\alpha)}{s}\right)^k\frac{a^s}{s!}}{1-\frac{a(1-q-\alpha)}{s}}\frac{p_{0,0}}{1-q\frac{a^s}{s!}\left(\frac{a(1-\alpha)}{s}\right)^k\frac{1}{1-\frac{a(1-q-\alpha)}{s}}},$$

$$P_b = \frac{\alpha\frac{a^s}{s!}p_{0,0}\left(\frac{1-\left(\frac{a(1-\alpha)}{s}\right)^k}{1-\frac{a(1-\alpha)}{s}} + \frac{\left(\frac{a(1-\alpha)}{s}\right)^k}{1-\frac{a(1-q-\alpha)}{s}}\right)}{1-q\frac{a^s}{s!}\left(\frac{a(1-\alpha)}{s}\right)^k\frac{1}{1-\frac{a(1-q-\alpha)}{s}}},$$

$$E(W_1) = \frac{\frac{a^s}{s!}}{\lambda(1-\Psi-P_b)} \frac{p_{0,0}\left(\sum_{x=0}^{k-1} x\left(\frac{a(1-\alpha)}{s}\right)^x + \left(\frac{a(1-\alpha)}{s}\right)^k \left(\frac{k\left(1-\frac{a(1-q-\alpha)}{s}\right)+\frac{a(1-q-\alpha)}{s}}{\left(1-\frac{a(1-q-\alpha)}{s}\right)^2}\right)\right)}{1-q\frac{a^s}{s!}\left(\frac{a(1-\alpha)}{s}\right)^k \frac{1}{1-\frac{a(1-q-\alpha)}{s}}},$$

with

$$p_{0,0} = \left[1 + \frac{\left(\sum_{x=0}^{s-2}\frac{a^{x+1}}{(x+1)!} + \frac{a^s}{s!}\sum_{x=0}^{k-1}\frac{(a(1-\alpha))^x}{s^x} + \frac{a^s}{s!}\frac{(a(1-\alpha))^k}{s^k}\frac{1}{1-\frac{a(1-q-\alpha)}{s}}\right)}{1-q\frac{a^s}{s!}\frac{(a(1-\alpha))^k}{s^k}\frac{1}{1-\frac{a(1-q-\alpha)}{s}}}\right]^{-1}.$$

We now consider the non-idling policy with $y_0 = s$. We have

$$\Psi = \frac{q\left(\frac{a(1-\alpha)}{s}\right)^k \frac{a^s}{s!}p_{0,0}}{1-\frac{a(1-q-\alpha)}{s} - q\frac{a}{s}\left(\frac{a(1-\alpha)}{s}\right)^k},$$

$$P_b = \frac{\alpha\frac{a^s}{s!}p_{0,0}}{1-\frac{(a(1-\alpha))}{s}},$$

$$E(W_1) = \frac{1}{\lambda}\frac{\frac{a^s}{s!}p_{0,0}\left(\left(\frac{a(1-\alpha)}{s}\right)\frac{1-k\left(\frac{a(1-\alpha)}{s}\right)^{k-1}+(k-1)\left(\frac{a(1-\alpha)}{s}\right)^k}{\left(1-\frac{a(1-\alpha)}{s}\right)^2} + \left(\frac{a(1-\alpha)}{s}\right)^k \frac{\frac{(1-q-\alpha)a}{s}+k\left(1-\frac{(1-q-\alpha)a}{s}\right)}{\left(1-\frac{(1-q-\alpha)a}{s}\right)^2}\right)}{\left(1-\frac{q\frac{a}{s}\left(\frac{a(1-\alpha)}{s}\right)^k}{1-\frac{a(1-q-\alpha)}{s}}\right)(1-\Psi-P_b)},$$

with

$$p_{0,0} = \left[\sum_{x=0}^{s-1}\frac{a^x}{x!} + \frac{\frac{a^s}{s!}}{1-\frac{(a(1-\alpha))}{s}}\right]^{-1}.$$

Using the fact that the overall system is equivalent to an M/M/s queue with balking, we can also compute $E(W_2)$ in the non-idling case as follows.

$$\begin{aligned} E(W) &= \Psi E(W_2) + (1-\Psi-P_b)E(W_1) \\ &= \frac{p_{0,0}\frac{a^s}{s!}}{\lambda}\frac{\frac{a(1-\alpha)}{s}}{\left(1-\frac{a(1-\alpha)}{s}\right)^2}. \end{aligned}$$

# E Proof of Proposition 4

Since $E(W_1)$ and $E(W_2)$ are both increasing in $k$, we choose the optimal value of $k$ which is $k = 0$. We rewrite the performance measures as functions of the Erlang Delay Loss Formulae, $C_s$ and the parameter $\rho = \frac{\lambda}{s\mu}$. Under the stability constraint $\lambda(1-\alpha) < s\mu$, we have

$$C_s = \frac{1}{1+s!\sum_{x=0}^{s-1}\frac{(s\rho)^{x-s}}{x!}(1-\rho(1-\alpha))}, \Psi = qC_s, P_b = \alpha C_s, E(W_1) = \frac{1}{\lambda q}\frac{\Psi\left(\frac{(1-q-\alpha)\rho}{1-(1-q-\alpha)\rho}\right)}{1-\Psi-P_b},$$

and

$$E(W) = \frac{\Psi\rho(1-\alpha)}{\lambda q\left(1-\rho(1-\alpha)\right)}.$$

Harel (2011) shows that the Erlang loss formulae is strictly decreasing in $s$. In particular, he shows that the function $\varphi(s) = s!\sum\limits_{x=0}^{s-1}\frac{(s\rho)^{x-s}}{x!}$ is strictly increasing in $s$. This expression is in the denominator of $C_s$ and is multiplied by the positive coefficient $1-\rho(1-\alpha)$. We therefore deduce that $C_s$ is decreasing in $s$ as well as $\Psi$, $P_b$ and $E(W)$, because they are all proportional to $C_s$. Since $\Psi$ and $P_b$ are decreasing in $s$, $1-P_b-\Psi$ is increasing in $s$ and $E(W_1)$ is decreasing in $s$.

Harel (2011) also shows that $\frac{1}{1+\varphi(s)}$ is strictly convex in $s$ (convexity of the Erlang loss formula) and that $\frac{1}{1+s!\sum\limits_{x=0}^{s-1}\frac{(s\rho)^{x-s}}{x!}(1-\rho)} = \frac{1}{1+\varphi(s)(1-\rho)}$ is strictly convex in $s$ (convexity of the Erlang delay formula). Since $C_s = \frac{1}{1+\varphi(s)(1-\rho(1-\alpha))}$, one may write

$$\frac{\partial^2 C_s}{\partial s^2} = (1-\rho(1-\alpha))\frac{2(1-\rho(1-\alpha))(\varphi'(s))^2-\varphi''(s)(1+\varphi(s))}{\left(1+\varphi(s)(1-\rho(1-\alpha))\right)^3}.$$

From Harel (2011), we have $2(1-\rho(1-\alpha))(\varphi'(s))^2-\varphi''(s)(1+\varphi(s)) > 0$ for $\alpha = 0$ and $\alpha = 1$. Since $\varphi(s)$ does not depend on $\alpha$ and $1-\rho(1-\alpha)$ is strictly increasing in $\alpha$, we obtain $\frac{\partial^2 C_s}{\partial s^2} > 0$. Therefore, $C_s$ is strictly convex in $s$.

We next deduce that $\Psi$, $P_b$ and $E(W)$ are convex in $s$. One may see that $E(W_1)$ is proportional to $D_s$, with $D_s = \frac{C_s}{1-(\alpha+q)C_s}$. We have

$$\begin{aligned} D_{s+2}+D_s-2D_{s+1} &= \frac{C_{s+2}+C_s-2C_{s+1}+(\alpha+q)\left(C_sC_{s+1}+C_{s+1}C_{s+2}-2C_sC_{s+2}\right)}{(1-(\alpha+q)C_s)(1-(\alpha+q)C_{s+1})(1-(\alpha+q)C_{s+2})} \\ &= \frac{(C_{s+2}+C_s-2C_{s+1})(1-(\alpha+q)C_{s+1})+2(\alpha+q)(C_s-C_{s+1})(C_{s+1}-C_{s+2})}{(1-(\alpha+q)C_s)(1-(\alpha+q)C_{s+1})(1-(\alpha+q)C_{s+2})}. \end{aligned}$$

Since $C_s$ is strictly convex, $C_{s+2}+C_s-2C_{s+1} > 0$. Since $C_s$ is strictly decreasing, $(C_s-C_{s+1})(C_{s+1}-C_{s+2}) > 0$. Thus, $D_s$ is strictly convex in $s$ and $E(W_1)$ is also strictly convex in $s$. This finishes the proof of the proposition. □

# F Proof of Proposition 5

Recall that in the non-idling case, we have

$$\Psi = \frac{q\left(\frac{a(1-\alpha)}{s}\right)^k\frac{a^s}{s!}p_{0,0}}{1-\frac{a(1-q-\alpha)}{s}-q\frac{a}{s}\left(\frac{a(1-\alpha)}{s}\right)^k}.$$

Since for stability reason $\frac{a(1-\alpha)}{s} \leq 0$, we obtain

$$\frac{\partial\Psi}{\partial k} = q\frac{a^s}{s!}p_{0,0}\frac{\left(\frac{a(1-\alpha)}{s}\right)^k\ln\left(\frac{a(1-\alpha)}{s}\right)\left(1-\frac{a(1-q-\alpha)}{s}\right)}{\left(1-\frac{a(1-q-\alpha)}{s}-q\frac{a}{s}\left(\frac{a(1-\alpha)}{s}\right)^k\right)^2} \leq 0.$$

Thus, $\Psi$ is decreasing in $k$. We now rewrite $E(W_1)$ as

$$E(W_1) = \frac{\left(1 - \frac{(1-q-\alpha)a}{s}\right)}{\lambda q} f(k) g(k),$$

where

$$f(k) = \frac{\Psi}{\left(\frac{a(1-\alpha)}{s}\right)^k (1 - \Psi - P_b)},$$

and

$$g(k) = \left( \left(\frac{a(1-\alpha)}{s}\right) \frac{1 - k\left(\frac{a(1-\alpha)}{s}\right)^{k-1} + (k-1)\left(\frac{a(1-\alpha)}{s}\right)^k}{\left(1 - \frac{a(1-\alpha)}{s}\right)^2} + \left(\frac{a(1-\alpha)}{s}\right)^k \frac{\frac{(1-q-\alpha)a}{s} + k\left(1 - \frac{(1-q-\alpha)a}{s}\right)}{\left(1 - \frac{(1-q-\alpha)a}{s}\right)^2} \right).$$

First we show that $f(k)$ is increasing in $k$. We have

$$f'(k) = \left(\frac{a(1-\alpha)}{s}\right)^k \frac{-\frac{\partial \Psi}{\partial k} P_b - \Psi \ln\left(\frac{a(1-\alpha)}{s}\right)(1 - \Psi - P_b)}{\left(\left(\frac{a(1-\alpha)}{s}\right)^k (1 - \Psi - P_b)\right)^2} \geq 0,$$

because $\Psi$ is decreasing in $k$ and $\frac{a(1-\alpha)}{s} < 1$ for stability reason. We rewrite $g(k)$ as

$$g(k) = \frac{\frac{a(1-\alpha)}{s}\left(1 - \frac{(1-q-\alpha)a}{s}\right)^2 - \frac{qa}{s}\left(1 - \frac{a(1-\alpha)}{s}\right)^k \left(k\left(1 - \frac{a(1-\alpha)}{s}\right)\left(1 - \frac{(1-q-\alpha)a}{s}\right) + 1 - \frac{a(1-\alpha)}{s}\frac{(1-q-\alpha)a}{s}\right)}{\left(1 - \frac{a(1-\alpha)}{s}\right)^2 \left(1 - \frac{(1-q-\alpha)a}{s}\right)^2}.$$

Only the numerator, say $n(k)$, of this expression depends on $k$. We have

$$n'(k) = -\frac{qa}{s} \ln\left(\frac{a(1-\alpha)}{s}\right)\left(1 - \frac{a(1-\alpha)}{s}\right)^k \left((k+1)\left(1 - \frac{a(1-\alpha)}{s}\right)\left(1 - \frac{(1-q-\alpha)a}{s}\right) + 1 - \frac{a(1-\alpha)}{s}\frac{(1-q-\alpha)a}{s}\right) \geq 0,$$

since $\frac{a(1-\alpha)}{s} < 1$. We finally deduce that $E(W_1)$ is increasing in $k$, which completes the proof of the proposition. □

# G Proof of Proposition 6

When an idle agent considers the service of the first outbound call in line, There are two possibilities. The first possibility (with probability $r_1 + r_2 > 0$) is that the customer is available and will be served within an exponential duration with parameter $\mu_1$ or $\mu_2$ with probability $r_1$ or $r_2$, respectively. The second possibility (with probability $1 - r_1 - r_2$) is that the customer is not available and the agent will be occupied a random duration exponentially distributed with parameter $\mu_3$. This customer will be then called back again latter according to the same process and independently of the fact that she has been already called back. Let us denote by $U_i$, a Bernouilli random variable, which takes the value 1 with probability $r_1 + r_2$ and 0 otherwise for $i \geq 1$; by $V_i$, a Bernouilli random variable, which takes the value 1 with probability $\frac{r_1}{r_1+r_2}$ and 0 otherwise for $i \geq 1$; and by $T_{i,j}$ an exponential random variable with parameter $\mu_j$, for $i \geq 1$ and $j = 1, 2, 3$. The time duration, denoted by the random variable $T$, which is spent by the system capacity to serve an outbound call, can be written as follows.

$$\begin{aligned} T &= U_1(V_1 T_{1,1} + (1 - V_1)T_{1,2}) + (1 - U_1)(T_{1,3} + U_2(V_2 T_{2,1} + (1 - V_2)T_{2,2}) + (1 - U_2)(T_{2,3} + \cdots \\ &= \sum_{i=1}^{\infty} \prod_{k=1}^{i-1} (1 - U_k) U_i V_i T_{i,1} + \sum_{i=1}^{\infty} \prod_{k=1}^{i-1} (1 - U_k) U_i (1 - V_i) T_{i,2} + \sum_{i=1}^{\infty} \prod_{k=1}^{i} (1 - U_k) T_{i,3}. \end{aligned}$$

We next derive the expected value of $T$. Since all the considered random variables are independent, we have

$$\begin{aligned}E(T) &= \sum_{i=1}^{\infty}\prod_{k=1}^{i-1} E(1-U_k)E(U_i)E(V_i)E(T_{i,1}) + \sum_{i=1}^{\infty}\prod_{k=1}^{i-1} E(1-U_k)E(U_i)E(1-V_i)E(T_{i,2}) + \sum_{i=1}^{\infty}\prod_{k=1}^{i} E(1-U_k)E(T_{i,3})\\ &= \frac{r_1}{r_1+r_2}\frac{1}{\mu_1} + \frac{r_2}{r_1+r_2}\frac{1}{\mu_2} + \frac{1-r_1-r_2}{r_1+r_2}\frac{1}{\mu_3}.\end{aligned}$$

We now derive the variance of $T$, denoted by $Var(T)$. Again, from the independence of the random variables, we obtain

$$Var(T) = \sum_{i=1}^{\infty} Var(\prod_{k=1}^{i-1}(1-U_k)U_iV_iT_{i,1}) + \sum_{i=1}^{\infty} Var(\prod_{k=1}^{i-1}(1-U_k)U_i(1-V_i)T_{i,2}) + \sum_{i=1}^{\infty} Var(\prod_{k=1}^{i}(1-U_k)T_{i,3}).$$

Let us define the sequence $S_n$ by $S_n = Var\left(\prod_{k=1}^{n}(1-U_k)\right)$, for $n \geq 0$, with $S_0 = 0$. We have

$$S_n = S_{n-1}(Var(1-U_n) + E^2(1-U_n)) + Var(1-U_n)E^2\left(\prod_{k=1}^{n-1}(1-U_k)\right),$$

for $n \geq 1$. Since $Var(1-U_n) = (r_1+r_2)(1-r_1-r_2)$, $E^2(1-U_n) = (1-r_1-r_2)^2$ and $E^2\left(\prod_{k=1}^{n-1}(1-U_k)\right) = (1-r_1-r_2)^{2n-2}$, we obtain

$$S_n = (1-r_1-r_2)S_{n-1} + (r_1+r_2)(1-r_1-r_2)^{2n-1}, \tag{64}$$

for $n \geq 1$. Using Equation (64), it is easy to prove by induction that $S_n = (1-r_1-r_2)^n(1-(1-r_1-r_2)^n)$, for $n \geq 0$. We next compute $Var(U_nV_nT_{n,1})$, for $n \geq 1$. We may write

$$\begin{aligned}Var(U_nV_nT_{n,1}) &= Var(U_nV_n)Var(T_{n,1}) + Var(U_nV_n)E^2(T_{n,1}) + E^2(U_nV_n)Var(T_{n,1})\\ &= \frac{1}{\mu_1^2}\left(2Var(U_nV_n) + E^2(U_nV_n)\right)\\ &= \frac{1}{\mu_1^2}\left(2Var(U_n)Var(V_n) + 2E^2(U_n)Var(V_n) + 2Var(U_n)E^2(V_n) + E^2(U_n)E^2(V_n)\right)\\ &= \frac{r_1}{\mu_1^2}\left(\frac{2(1-r_1-r_2)r_2}{r_1+r_2} + 2r_2 + 2\frac{(1-r_1-r_2)r_1}{r_1+r_2} + r_1\right)\\ &= \frac{r_1(2-r_1)}{\mu_1^2}.\end{aligned}$$

Hence

$$\begin{aligned}&\sum_{i=1}^{\infty} Var(\prod_{k=1}^{i-1}(1-U_k)U_iV_iT_{i,1})) = \sum_{i=1}^{\infty}\left(S_{i-1}(Var(U_iV_iT_{i,1}) + E^2(U_iV_iT_{i,1})) + E^2(\prod_{k=1}^{i-1}(1-U_k))Var(U_iV_iT_{i,1})\right)\\ &= \sum_{i=1}^{\infty}\left((1-r_1-r_2)^{i-1}(1-(1-r_1-r_2)^{i-1})(\frac{r_1(2-r_1)}{\mu_1^2} + \frac{r_1^2}{\mu_1^2}) + (1-r_1-r_2)^{2i-2}\frac{r_1(2-r_1)}{\mu_1^2}\right)\\ &= \frac{2r_1}{\mu_1^2}\sum_{i=1}^{\infty}(1-r_1-r_2)^{i-1} - \frac{r_1^2}{\mu_1^2}\sum_{i=1}^{\infty}(1-r_1-r_2)^{2(i-1)}\\ &= \frac{r_1(4-3r_1-2r_2)}{\mu_1^2(r_1+r_2)(2-r_1-r_2)}.\end{aligned}$$

Using the same approach, we also obtain $\sum_{i=1}^{\infty} Var(\prod_{k=1}^{i-1}(1-U_k)U_i(1-V_i)T_{i,2})) = \frac{r_2(4-3r_2-2r_1)}{\mu_2^2(r_1+r_2)(2-r_1-r_2)}$ and

$\sum_{i=1}^{\infty} Var(\prod_{k=1}^{i}(1-U_k)T_{i,3}) = \frac{(1-r_1-r_2)(4-r_1-r_2)}{\mu_3^2(r_1+r_2)(2-r_1-r_2)}$. This finishes the proof of the proposition. □